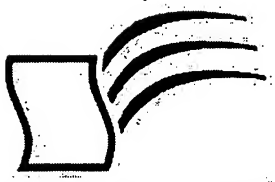
Christian Hennig

Datenanalyse mit Modellen für Cluster linearer Regression

Dissertation
an der Universität Hamburg
Fachbereich Mathematik
Institut für Mathematische Stochastik
Mai 1997 Abgabe



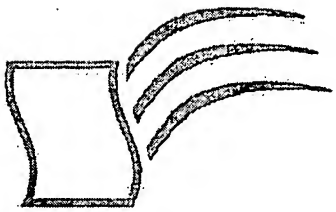
Diplomarbeiten Agentur
Dipl. Kfm. Dipl. Hdl. Björn Bedey
Dipl. Wi.-Ing. Martin Haschke
und Guido Meyer GbR

Hermannstal 119 k
22119 Hamburg

agentur@diplom.de
www.diplom.de

Hennig, Christian: Datenanalyse mit Modellen für Cluster linearer Regression /
Christian Hennig - Hamburg: Diplomarbeiten Agentur, 2000
Zugl.: Hamburg, Universität, Dissertation, 1997

Dipl. Kfm. Dipl. Hdl. Björn Bedey, Dipl. Wi.-Ing. Martin Haschke & Guido Meyer GbR
Diplomarbeiten Agentur, http://www.diplom.de, Hamburg 1999
Printed in Germany



Diplomarbeiten Agentur

Wissensquellen gewinnbringend nutzen

Qualität, Praxisrelevanz und Aktualität zeichnen unsere Studien aus. Wir bieten Ihnen im Auftrag unserer Autorinnen und Autoren Wirtschaftsstudien und wissenschaftliche Abschlussarbeiten – Dissertationen, Diplomarbeiten, Magisterarbeiten, Staatsexamensarbeiten und Studienarbeiten zum Kauf. Sie wurden an deutschen Universitäten, Fachhochschulen, Akademien oder vergleichbaren Institutionen der Europäischen Union geschrieben. Der Notendurchschnitt liegt bei 1,5.

Wettbewerbsvorteile verschaffen – Vergleichen Sie den Preis unserer Studien mit den Honoraren externer Berater. Um dieses Wissen selbst zusammenzutragen, müssten Sie viel Zeit und Geld aufbringen.

http://www.diplom.de bietet Ihnen unser vollständiges Lieferprogramm mit mehreren tausend Studien im Internet. Neben dem Online-Katalog und der Online-Suchmaschine für Ihre Recherche steht Ihnen auch eine Online-Bestellfunktion zur Verfügung. Inhaltliche Zusammenfassungen und Inhaltsverzeichnisse zu jeder Studie sind im Internet einsehbar.

Individueller Service – Gerne senden wir Ihnen auch unseren Papierkatalog zu. Bitte fordern Sie Ihr individuelles Exemplar bei uns an. Für Fragen, Anregungen und individuelle Anfragen stehen wir Ihnen gerne zur Verfügung. Wir freuen uns auf eine gute Zusammenarbeit

Ihr Team der *Diplomarbeiten* Agentur

Dipl. Kfm. Dipl. Hdl. Björn Bedey –
Dipl. Wi.-Ing. Martin Haschke –
und Guido Meyer GbR

Hermannstal 119 k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

agentur@diplom.de
www.diplom.de

Als Dissertation angenommen vom Fachbereich Mathematik der Universität Hamburg

auf Grund der Gutachten von Prof. Dr. Konrad Behnen
und Prof. Dr. Dietmar Pfeifer.

Hamburg, den 16.5.1997

Prof. Dr. Reiner Hass
Sprecher des Fachbereichs Mathematik

Diese Arbeit ist Prof. Gerhard „Heik" Portele vom Zentrum für Hochschuldidaktik gewidmet. Er hat mir nahegebracht und vorgelebt, daß es beim Lehren in der Hochschule möglich ist, Autonomie zu gewähren, auf Machtausübung zu verzichten und in lebendigem Kontakt mit den Lehrenden zu bleiben. Prof. Portele starb am 10.7.1996 an Krebs.

Danksagung: Mein größtes Dankeschön gilt meinem Betreuer Prof. Konrad Behnen für Aufgeschlossenheit und Anregungen sowie für seine kritischen Anmerkungen über Darstellung und Lesbarkeit. Von diesen Hinweisen werden vermutlich die Leser/innen nicht nur dieser Arbeit sehr profitieren. Prof. Dietmar Pfeifer danke ich, daß er die Arbeit des Zweitgutachters übernommen hat. Ein weiterer besonderer Dank geht an Vanessa Didelez, Gabi Schneider, Dr. Silvelyn Zwanzig und Dr. Lutz Mattner, die Teile der Arbeit durchgesehen und wertvolle Kommentare zu Form und Inhalt gegeben haben. Zuletzt möchte ich allen weiteren Menschen danken, die sich für meine Arbeit interessiert, mir Literatur- und andere sinnvolle Hinweise gegeben oder mich in anderer Weise unterstützt haben.

Abstract[1]

A linear regression can be modeled by a family of distributions $(P_{\beta,\sigma^2} : \beta \in I\!R^{p+1}, \sigma^2 \in I\!R^+)$ for $(x, y) \in I\!R^{p+1} \times I\!R$, where $y = x'\beta + u$, u independent of x and distributed normal or symmetrically about 0 with variance σ^2.

This thesis deals with the analysis of datasets $(x_i, y_i) \in I\!R^{p+1} \times I\!R, i = 1, \ldots, n$. A linear regression distribution P_{β,σ^2} is treated as a distribution for one cluster, i.e. linear regression distributions with different parameters $(\beta_i, \sigma_i^2), i = 1, \ldots, s$ are supposed to be adequat for different parts of the dataset. Furthermore, there can be outliers in the data for which no such model is appropriate.

Various models for such data are introduced, especially mixture models of the form $\sum_{i=1}^{s} \epsilon_i P_{\beta_i,\sigma_i^2}$. Maximum Likelihood estimation of the parameters (β_i, σ_i^2) is discussed. New proposals for estimating the number of clusters s are given.

Sufficient conditions for the identifiability of the parameters are derived. Counter-examples are given in some situations where the conditions do not hold.

As a new method, Fixed Point Cluster Analysis (FPCA) is introduced. It enables the analysis of data with unknown number of clusters s and outliers. FPCA bases on the identification of outliers and can be generalized to other clustering problems. A Fixed Point Cluster (FPC) corresponds to a subset of $I\!R^{p+1} \times I\!R$ and should contain points (x, y) which belong together in some sense. Every FPC corresponds to parameters $(b, s^2) \in I\!R^{p+1} \times I\!R^+$ which can be interpreted as estimation of the regression parameters (β_i, σ_i^2). FPC are defined for datasets and distributions.

Convergence of an algorithm for the computation of FPC for given datasets is proven.

Distributions of the form $(1-\epsilon)P_{\beta_0,\sigma_0^2} + \epsilon H^*$ are considered. P_{β_0,σ_0^2} here is interpreted as a distribution for a linear regression cluster. H^* is a distribution on $I\!R^{p+1} \times I\!R$, e.g. a mixture of other P_{β,σ^2}. The existence of FPC is shown under various assumptions on H^* and ϵ. The parameters of these FPC lie in a bounded neighborhood of (β_0, σ_0^2). For homogenous regression distributions ($\epsilon = 0$) exists one and only one FPC. It has parameters (β_0, σ_0^2).

In a simulation study FPCA and two Maximum Likelihood procedures are compared.

[1]Eine deutsche Zusammenfassung findet sich auf Seite 179.

Inhaltsverzeichnis

1 Einführung

1.1 Das Problem

Der folgende Datensatz findet sich auf Seite 26 von Rousseeuw und Leroy (1988). Er enthält die von Belgien aus geführten internationalen Telefongespräche (in 10 Millionen) in den Jahren 1950-1973.

Nr.	Telefonate (y)	Jahr (x)		Nr.	Telefonate (y)	Jahr (x)
1	0.44	50		13	1.61	62
2	0.47	51		14	2.12	63
3	0.47	52		15	11.9	64
4	0.59	53		16	12.4	65
5	0.66	54		17	14.2	66
6	0.73	55		18	15.9	67
7	0.81	56		19	18.2	68
8	0.88	57		20	21.2	69
9	1.06	58		21	4.3	70
10	1.20	59		22	2.4	71
11	1.35	60		23	2.7	72
12	1.49	61		24	2.9	73

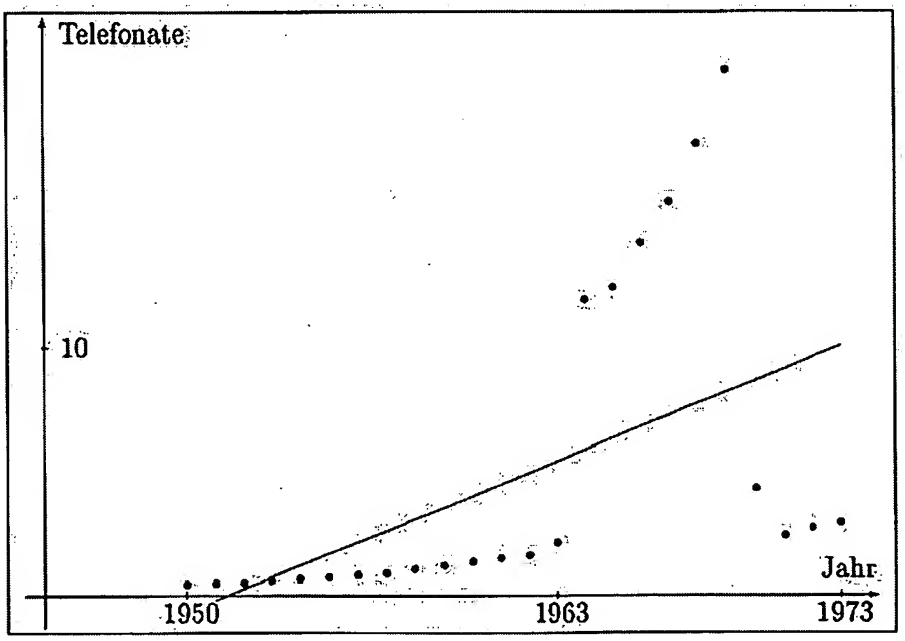


Abbildung 1: Telefondatensatz

In Abbildung 1 fällt sofort auf, daß sich die Telefonate in den Jahren von 1964-1970 grundsätzlich anders verhalten als die Mehrheit der Daten. Der Zusammenhang zwischen Jahr und Telefonatezahl sieht für die Jahre 1950-1962 und 1971-1973 annähernd linear

aus. Auf Nachfrage erfuhren Rousseeuw und Leroy, daß 1964-1969 nicht die Telefonate, sondern die Minuten gezählt wurden, die die Telefonate insgesamt dauerten. 1963 und 1970 wurden beide Verfahren teilweise angewendet.

In der robusten Statistik wurde dieser Datensatz häufig diskutiert als Beispiel für eine lineare Regression mit mehreren Ausreißern. Berechnet man den Kleinste-Quadrate-Schätzer (KQ) zum Modell

$$y = \beta_1 x + \beta_2 + u,\ E(u) = 0,$$

so paßt die resultierende Gerade fast keinen Punkt gut an (steigende Linie in der Abbildung 1).

Bemerkung 1.1 *Gegeben sei ein Datensatz* $(\mathbf{X}, y)$,

$$\mathbf{X} = (x_1, \ldots, x_n)',\ y = (y_1, \ldots, y_n)',\ x_i \in I\!R^{p+1},\ y_i \in I\!R\ \forall i = 1, \ldots, n.$$

Dann ist der KQ-Schätzer $\hat{\beta}_{KQ} \in I\!R^{p+1}$ *definiert durch*

$$\sum_{i=1}^{n} (y_i - x_i' \hat{\beta}_{KQ})^2 \stackrel{!}{=} \min.$$

Das heißt, falls $(\mathbf{X}'\mathbf{X})^{-1}$ *existiert,* $\hat{\beta}_{KQ} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$. *Im obigen Fall der Regression mit Achsenabschnitt* β_2 *werden die* $x_i = (x_{i1}, x_{i2})$ *als Elemente aus* $I\!R^2$ *interpretiert, wobei immer* $x_{i2} = 1$.

Wählt man aber einen robusten Regressionsschätzer wie zum Beispiel „Least Median of Squares" (siehe Rousseeuw und Leroy (1988)), so wird eine Gerade geschätzt, die nur die Mehrheit der Jahre gut anpaßt, in denen die Gespräche gezählt wurden. Die Daten werden also unterteilt in „gute" Daten und „Ausreißer". Aber was ist mit den Daten von 1964-69? Sie sind ja nicht falsch, sondern nur andersartig. Besteht bei ihnen vielleicht auch ein einfacher linearer Zusammenhang? Da es so wenige sind, ist das vom optischen Eindruck her nicht klar. Inhaltlich wird man zumindest einen approximativ linearen Zusammenhang bei den Gesprächslängen vermuten, falls Linearität für die Anrufsanzahlen vorausgesetzt wird.

Das Thema dieser Arbeit ist die Clusteranalyse von Daten aus linearen Regressionen. Das heißt: Es geht darum, Gruppen von Daten zu finden, wobei Daten zusammen eine Gruppe bilden sollen, wenn sie durch denselben linearen Zusammenhang zwischen der (möglicherweise mehrdimensionalen) Regressorvariable x und der (eindimensionalen) abhängigen Variablen y erzeugt wurden. Zur Modellierung der Daten einer Gruppe soll also ein klassisches lineares Regressionsmodell (siehe (2.1) in Abschnitt 2) adäquat sein. Zu beachten ist dabei, daß hier im Unterschied zur häufigsten Verwendung des Wortes „Cluster" (Klumpen) die Zusammengehörigkeit von Punkten nicht direkt mit ihrem Abstand voneinander zusammenhängt. Das ist in Abbildung 1 zum Beispiel zu sehen, wenn man den Punkt für 1973 betrachtet, der weiter vom 1950er Punkt entfernt ist als von sämtlichen „Ausreißern".

Zu diesem Ziel werden zunächst Maximum Likelihood- und andere bekannte Ansätze untersucht. Dann führe ich im Hauptteil der Arbeit die Fixpunktclusteranalyse ein, die speziell zur Clusteranalyse bei Clustern unterschiedlicher Art und Präsenz von Ausreißern dienen soll.

1.2 Modelle für die Clusteranalyse (Teil I)

Der übliche stochastische Zugang zu einem Clusteranalyse-Problem ist die Formulierung eines möglichst einfachen Clustermodelles[2]. Innerhalb dieses Modells kann dann nach Schätzern mit guten Eigenschaften für die Regressions- und Störskalenparameter der einzelnen Cluster gesucht werden.

Es gibt zwei unterschiedliche Methoden, stochastische Modelle für die Clusteranalyse zu formulieren: Mischmodelle, d.h. Modelle, bei denen die Punkte unabhängig identisch verteilt sind. Die Werte werden mit festgelegten, aber unbekannten Wahrscheinlichkeiten aus unterschiedlichen Populationen erzeugt. In Modellen mit fester Zuordnung sind die Punkte unterschiedlicher Cluster dagegen unterschiedlich verteilt und die Zugehörigkeit eines Punktes zu einem Cluster wird als fester, unbekannter Modellparameter behandelt. In Abschnitt 2 werden die unterschiedlichen Modelle vorgestellt.

Ein Spezialfall der zweiten Modellvariante sind Wechselpunktprobleme („changepoint problems"), über die im Regressionsfall am meisten bekannt ist. In einem Wechselpunktmodell ändern sich die Regressionsparameter in Abhängigkeit von der Zeit oder anderen Regressoren. In Abschnitt 3.1 wird ein kurzer Überblick über die diesbezügliche Literatur gegeben. Ein solches Modell könnte auch für den Telefondatensatz benutzt werden. Allerdings wird in der Literatur über Wechselpunktprobleme normalerweise nicht vorgesehen, daß ein System wieder in den alten Zustand zurückspringt (im Datensatz nach 1970).

Weiter wurden Kleinste-Quadrate- und Maximum Likelihood(ML)-Schätzer für den Fall vorgeschlagen, daß die Zugehörigkeit der Punkte zu den Clustern als unabhängig von den Regressoren vorausgesetzt wird. Diese Ansätze werden in den Abschnitten 3.2 und 3.3 diskutiert. Über die theoretischen Eigenschaften dieser Schätzer gibt es bislang im Regressionsfall kaum wesentliche Resultate. Ein großer Teil der Literatur befaßt sich mit der Entwicklung konvergenter Algorithmen zur Berechnung der Schätzer. Für die Schätzung der Clusterzahl wird häufig die Minimierung von informationsbasierten Kriterien vorgeschlagen, für die es aber nur wenig theoretische Rechtfertigung gibt.

Im allgemeinen kann über Abhängigkeiten zwischen Regressoren und Regressionsparametern keine einfache Voraussetzung gemacht werden. Clustermodelle mit fester Zuordnung ohne die restriktiven Voraussetzungen des Wechselpunktproblems wurden bislang nur im Lokationsproblem[3] behandelt. In Abschnitt 3.4 übertrage ich einen ML-Ansatz von Scott und Symons (1971) auf den linearen Regressionsfall. Abschnitt 3.5 stellt kurz alternative Ansätze zur Behandlung des Regressions-Clusterproblemes vor. In den Teilen von Abschnitt 3 wird ein Überblick über die bisher vorhandene Literatur zur Problemstellung gegeben.

Eine wesentliche Voraussetzung für Resultate über konsistente Schätzungen in clustererzeugenden Modellen ist die Identifizierbarkeit der Modellparameter: Die Parameterwerte, die eine bestimmte Verteilung definieren, müssen eindeutig sein. In den Abschnitten 4 bis 6 wird die Identifizierbarkeit der vorgestellten Modelle untersucht.

---

[2]Unter einem „Modell" verstehe ich eine Familie von Verteilungen $\{P_\theta, \theta \in \Theta\}$ auf einem Raum mit einer σ–Algebra, üblicherweise $(\mathbb{R}^d, \mathcal{B}^d)$. Mit „Verteilung" meine ich P_θ für ein bestimmtes θ.

[3]Wenn in dieser Arbeit vom Lokationsproblem die Rede ist, dann ist die Analyse von Daten mit Modellen gemeint, in denen unterschiedliche Teilmengen der Datenpunkte (Cluster) durch Verteilungen der Form $F[\mathrm{A}(y-b)], y \in \mathbb{R}^p$ mit unterschiedlichen Parametern $b \in \mathbb{R}^p$ (Lokations-, Lageparameter) beschrieben werden sollen. Der Modellparameter $\mathrm{A} \in \mathbb{R}^{p^2}$ kann fest, frei, bekannt oder unbekannt sein.

Es stellt sich heraus, daß häufig nicht alle Parameter identifizierbar sind. Zum Beispiel sind im Modell mit fester Zuordnung die Zuordnungsparameter nicht identifizierbar. Für die Identifizierbarkeit von Regressions- und Störskalenparametern werden hinreichende Bedingungen an die Regressoren hergeleitet.

Der Telefondatensatz wirft aber Probleme auf, die mit der skizzierten Herangehensweise schwerlich zu lösen sind:

- Es ist nicht klar, ob ein Modell mit mehreren Clustern dem Datensatz angemessener ist als ein Modell mit einer Mehrheit von Daten aus demselben Regressionsmodell und einer Minderheit nicht näher spezifizierter Ausreißer.
- Es ist nicht klar, ob der Zusammenhang in allen Clustern linear ist.
- Es ist nicht klar, ob es Punkte gibt, die sinnvollerweise zu gar keinem oder mehreren Clustern dazugerechnet werden sollten. Was ist mit den Jahren 1963 und 1970, als die Zählung umgestellt wurde?

Diese Probleme tauchen nicht nur im Falle der Telefondaten auf. Welches Modell für einen gegebenen Datensatz angemessen ist, weiß man von vornherein nie.

1.3 Exkurs: Angemessenheit von Modellen

Für die Motivation der späteren Abschnitte spielt die Funktion von Modellen in der Datenanalyse eine große Rolle. Daher möchte ich kurz die Vorstellungen skizzieren, die für meine Arbeit maßgeblich sind.

Der Satz „Ein bestimmtes Modell ist angemessen für einen Datensatz" bedeutet sinnvollerweise nicht: „Der Datensatz ist von einer Verteilung dieses Modells generiert worden." Eine solche Aussage würde sich auf keine Weise verifizieren lassen, und es ist kaum vorstellbar, daß sie jemals stimmen könnte. Davies (1995) schreibt:

The term „adequate" reflects the philosophy that a model is not true nor even treated as true. The model is regarded as being adequate for some given purpose. (...) The adequacy region specifies those probability models whose samples typically look like the actual data.

Die „adequacy region" ist Davies' Ansatz, Angemessenheit formal zu definieren. „Typically look like" bedeutet hier, daß der Datensatz eine - je nach Interpretationsziel definierte - Eigenschaft hat, die Datensätze aus dem entsprechenden Modell mit hoher Wahrscheinlichkeit haben.

„Angemessenheit" hat bei Davies also zwei Aspekte:

- Erzeugt man künstlich Daten aus einer geeigneten Verteilung eines angemessenen Modells, so sollen diese Daten dem vorliegenden Datensatz ähnlich sehen.
- Der Begriff „ähnlich" ist subjektiv. Ob ein vorgegebener Datensatz einem typischen Modelldatensatz „ähnlich" sieht, hängt von Ähnlichkeitskriterien ab, die man selbst wählen muß.

Ein dritter wichtiger Aspekt ist, daß ein angemessenes Modell dafür geeignet sein sollte, die Fragen zu beantworten, die man an den Datensatz hat.

Zum Beispiel wäre ein homogenes lineares Regressionsmodell mit normalverteiltem Störterm u für den Telefondatensatz nicht angemessen: Die Residuen sind in auffälliger Weise und entgegen den Modellvoraussetzungen abhängig vom Regressor x (was zu formalisieren wäre, um Davies' Ansatz anzuwenden). Ein lineares Regressionsmodell für die Jahre 1950-1963 und 1971-1973 wird dagegen nach Davies' Kriterien kaum für unangemessen gehalten werden können. Es kann jedoch nicht alle Fragen an den Datensatz beantworten, wenn man sich dafür interessiert, wie die restlichen Jahren genau zu interpretieren sind. Eine Mischung aus zwei linearen Regressionsmodellen kann vermutlich die Jahre 1964 und 1970 nicht „angemessen" anpassen. In Abschnitt 10.1 wird der Telefon-Datensatz als Anwendungsbeispiel für die in dieser Arbeit betrachteten Verfahren diskutiert.

Es gibt auch Datensätze, die - bis auf die Verwendung eines Zufallszahlengenerators - tatsächlich aus einer Mischung mehrerer linearer Regressionen stammen, wobei aber diese Mischung mit statistischen Methoden kaum von einem geeigneten homogenen Modell oder einer Mischung mit ganz anderen Parameterwerten zu unterscheiden ist. Zum Beispiel ist x in Abbildung 2 verteilt nach $\mathcal{N}_{(0,1)}$, mit Wahrscheinlichkeit 0.5 ist $y = 0.5x + u$, mit derselben Wahrscheinlichkeit $y = -0.5x + u$, wobei u verteilt nach $\mathcal{N}_{(0,z)}$ erzeugt worden ist. Das erzeugende Modell ist dem Datensatz sicher nach Davies'schen Kriterien angemessen. Dennoch bringt die Analyse des Datensatzes mit den Regressionsparametern eines solchen Mischmodells offenbar keine anschaulich brauchbare, interpretierbare Vorstellung von den Daten.

Diese Diskussion soll verdeutlichen, was gemeint ist, wenn in dieser Arbeit von „Angemessenheit" die Rede ist. Das Wort wird allerdings informell benutzt. Das formale Konzept der „adequacy region" wird nicht weiter verwendet.

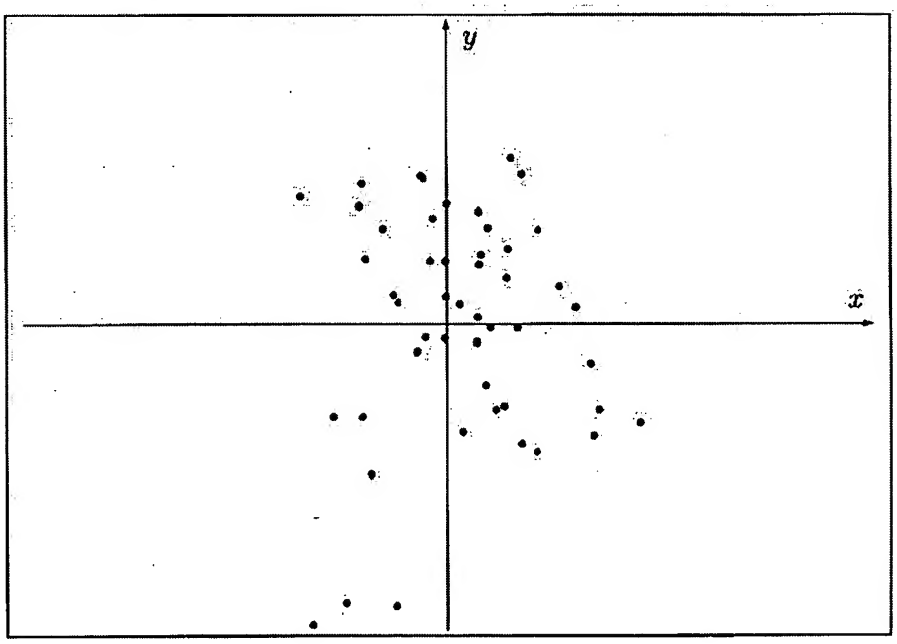


Abbildung 2: Gibt es hier Cluster?

1.4 Fixpunktcluster (Teil II und III)

Der Schwerpunkt dieser Arbeit liegt in der Entwicklung eines Clusteranalyse-Verfahrens („Fixpunktclusteranalyse"), das „anschaulich bedeutsame" Cluster finden soll. Dabei soll nicht vorausgesetzt werden müssen, daß eines der vorgestellten Modelle für den gesamten Datensatz paßt. Das Verfahren ist nur in dem Sinne modellbasiert, daß ein klassisches lineares Regressionsmodell für einen Teil der Daten angemessen sein soll, bzw. unterschiedliche Regressionsmodelle für unterschiedliche Teile der Daten. Diese Teile des Datensatzes sollen gefunden werden. Die Entwicklung des Verfahrens ist inspiriert durch die robuste Statistik und die Identifikation von Ausreißern. Abschnitt 3.5 diskutiert verwandte Ideen aus der robusten Statistik für die Clusteranalyse. Die Idee der Fixpunktcluster ist, daß keiner der Punkte eines Clusters bezüglich der für den Cluster angemessenen Verteilung Ausreißer ist, aber alle übrigen Punkte des Datensatzes. Ein Fixpunktcluster ist also ein Teildatensatz mit bestimmten Eigenschaften. Dadurch muß nicht jeder Punkt eines Datensatzes zu einem Cluster gehören, andererseits können sich die Cluster auch überschneiden. Mit dem oben diskutierten Begriff von „Angemessenheit" ist es denkbar, daß ein Punkt zu mehreren Teildatensätzen gehört, für die Regressionsmodelle mit unterschiedlichen Parametern angemessen sind.

Das Konzept wird in den Abschnitten 7.1 bis 8.2 eingeführt, nachdem der Begriff „Ausreißer" in Anlehnung an Davies und Gather (1993) formal definiert worden ist. Fixpunktcluster werden sowohl als exploratives Verfahren für Datensätze („Fixpunktclustervektoren") definiert, als auch als „Parameter" von stochastischen Modellen („Fixpunktclusterindikatoren"), als deren Schätzer das Datensatz-Verfahren dann interpretiert werden kann. Der Zusammenhang zwischen Fixpunktclustervektoren und -indikatoren kann bei geeigneter Definition darin bestehen, daß die Indikatoren Funktionale von W.-Maßen sind und man die Vektoren erhält, indem man diese Funktionale auf die empirischen Verteilungen anwendet. Die Anwendung der Fixpunktcluster-Idee ist nicht beschränkt auf das Regressionsproblem. In Abschnitt 7.2 wird anhand der Clusteranalyse von Datensätzen aus $(\{0,1\}^p)^n$ illustriert, wie Fixpunktcluster auch für andere Situationen definiert werden können. Im Regressionsfall ist jedem Fixpunktcluster ein Kennwert aus $\mathbb{R}^{p+1} \times \mathbb{R}^+$ zugeordnet, der als Schätzung des Regressionsparameters und der Störvarianz eines linearen Regressionsmodells betrachtet werden kann.

Das Fixpunktcluster-Verfahren beruht auf der Lösung einer Fixpunktgleichung. In Abschnitt 9 wird ein Algorithmus angegeben, um im Regressionsproblem in gegebenen Datensätzen Lösungen dieser Gleichung, d.h. Cluster, zu finden. Die Konvergenz dieses Algorithmus wird bewiesen.

Teil III ist den theoretischen Eigenschaften der Fixpunktcluster in stochastischen Modellen gewidmet. Die Resultate beziehen sich dabei nicht auf Datensätze, sondern auf das Verhalten von Fixpunktclusterindikatoren in den Modellen. Es wird bewiesen, daß es zu einer homogenen linearen Regressionsverteilung genau einen Fixpunktclusterindikator gibt. Dieser Fixpunktclusterindikator hat genau den Regressionsparameter und die Störvarianz der Verteilung als Kennwert. Weiterhin werden Modelle der Form $\epsilon H_0 + (1-\epsilon)H^*$, $0 < \epsilon \leq 1$ behandelt, wobei H_0 ein homogenes Regressionsmodell ist. In der robusten Statistik heißen diese Modelle „contamination model", da H_0 sozusagen durch H^* „verunreinigt" („contaminated") wird. Dieses sind die Modelle, für die die Fixpunktclusteranalyse angemessen sein sollte: Ein Teil der Daten folgt einem linearen Regressionsmodell, der Rest ist nicht näher spezifiziert. Mischmodelle haben zum Bei-

spiel diese Form, wobei dann H^* eine Mischung aus weiteren Regressionsmodellen wäre. Aufgrund rechnerischer Schwierigkeiten wird hierbei meistens ein eindimensionales Lokationsmodell (d.h. eine Regression, die nur aus einem Achsenabschnitt besteht) oder das (komplementäre) Regressionsmodell ohne Achsenabschnitt behandelt. Es wird unter gewissen Voraussetzungen an ϵ und H^* die Existenz von Fixpunktclusterindikatoren im „contamination model" bewiesen, deren Kennwerte von den Modellparametern von H_0 nur eine beschränkte (kleine) Abweichung haben.

Die Resultate sind mit Aussagen über „Fisher-Konsistenz" von Funktionalen vergleichbar. Daß die Regressionsparameter von Fixpunktclustervektoren in Datensätzen tatsächlich gegen Fixpunktclusterindikatoren konvergieren, falls $n \to \infty$, konnte ich bisher nicht zeigen. Dazu muß allerdings gesagt werden, daß vergleichbare (korrekte) Ergebnisse auch für die ML-Schätzer im Regressionsfall nicht existieren, für die solche Aussagen vermutlich leichter zu zeigen wären. Anders liegt die Situation dort allerdings bei anderen Problemstellungen, zum Beispiel im Lokationsfall. In Abschnitt 4 bei Bock (1996) wird hierzu ein Literaturüberblick gegeben. Bock weist dort auch darauf hin (Bemerkung 4.1), daß in der Clusteranalyse häufig die Konvergenz der Schätzer gegen ihre Funktionalwerte im Idealmodell geklärt ist, nicht aber die Entfernung dieser Funktionalwerte von den Modellparametern - im Gegensatz zur hier für Fixpunktcluster betriebenen Theorie.

Der praktische Wert der Theorie über Fixpunktclusterindikatoren ist schwer zu beurteilen. Es kann nicht quantifiziert werden, in welchem Maße die Ergebnisse für die Anwendung auf Datensätze relevant sind. Das ginge nur, wenn die Indikatoren schwach stetige Funktionale wären, d.h. wenn die Werte für empirische Verteilungen, die den theoretischen Modellen benachbart sind, in der Nähe der Modellindikatoren wären. Andererseits beleuchtet die Theorie den Zusammenhang zwischen dem heuristischen Konzept „Fixpunktcluster" und den Parametern clustererzeugender Modelle. Es zeigt sich zumindest in einigen Idealsituationen, daß ein „Fixpunktcluster" eine große Verwandtschaft zu den Komponenten von Mischverteilungen hat.

1.5 Vergleich der Verfahren (Teil IV)

Um die Wirkungsweise der Fixpunktclusteranalyse auf konkrete Datensätze zu erforschen und sie mit alternativen Möglichkeiten zu vergleichen, habe ich eine Simulationsstudie durchgeführt, die in Teil IV beschrieben wird. Die Simulationsstudie beschränkt sich auf Datensätze, die von einem der normalen clustergenerierenden Modelle aus Abschnitt 2 erzeugt wurden, also ohne nichtlineare Zusammenhänge und ohne Ausreißer, die die Modellvoraussetzungen verletzen. Weiterhin liegt ein Schwerpunkt auf Situationen mit gut voneinander getrennten Clustern. Im zweidimensionalen Fall heißt „gut getrennt" deutlich sichtbar. Diese Modelle sind sozusagen die „natürlichen Testobjekte" für die Verfahren. Datensätze mit „künstlichen Ausreißern" wurden erzeugt, indem ein sehr kleiner Anteil der Daten von linearen Regressionsverteilungen mit extrem hoher Störvarianz erzeugt wurde.

Als alternative Verfahren wurden die ML-Verfahren für das Mischmodell und das Modell mit fester Zuordnung verwendet. Es wurde immer davon ausgegangen, daß die Clusterzahl unbekannt ist.

Die Simulationen vermitteln einen Eindruck davon, wie schwierig die automatische

Behandlung der hier diskutierten Datenanalyse-Probleme ist. Für die Fixpunktclusteranalyse müssen offenbar Cluster wesentlich klarer voneinander getrennt sein als für die ML-Verfahren (und im niedrigdimensionalen Fall für das Auge), um sie zu finden. Das ist insofern kein Wunder, als daß die Clusterstruktur bei den ML-Verfahren Teil der Modellvoraussetzung ist. Andererseits haben die ML-Verfahren große Schwierigkeiten, wenn die Daten in der Nähe irregulärer Situationen sind, d.h. bei „ausreißererzeugenden Clustern" oder schwer identifizierbaren Parametern.

Meine Arbeit befindet sich im Grenzbereich zwischen mathematischer Statistik und Datenanalyse in einer Situation, die durch die schematische Anwendung von Standardverfahren nicht zu bewältigen ist. Neben der Diskussion solcher Standardverfahren und Standardmodelle habe ich mich bemüht, mit den Mitteln der statistischen Theorie und Simulation eine heuristische Idee zu untersuchen, die allenfalls teilweise modellbasiert ist. Die Arbeit stellt also eine neue Möglichkeit der Datenanalyse zur Diskussion. Neben ihrer kritischen Betrachtung werden Anregungen zur allgemeineren Verwendung und zur Verbesserung der Fixpunktclusteranalyse gemacht. Ich hoffe, daß einige der in dieser Arbeit angerissenen Ideen der weiteren Betrachtung würdig sind.

1.6 Formale und stilistische Bemerkungen

Nicht alle Teile der Arbeit bauen hierarchisch aufeinander auf. Die Arbeit hat sozusagen drei Richtungen, die weitgehend unabhängig voneinander lesbar sind:

- Modellierung und Identifizierbarkeit (Teil I, eventuell ohne Abschnitt 3),
- Theorie der Fixpunktcluster (Teil II und Teil III),
- Empirischer Vergleich von Verfahren zur Analyse von Mischungen linearer Regressionen (Abschnitte 2 und 3, Teil II und Teil IV).

Im Gegensatz zur Mehrheit der mathematischen Arbeiten taucht das Wort „ich" bei mir häufig auf. Arbeiten, die ausschließlich passiv formuliert sind, machen auf mich selten einen lebendigen Eindruck. Formulierungen mit „wir" suggerieren eine Übereinstimmung zwischen Autor/in und Leser/in, die nicht unbedingt vorhanden sein muß. Ich habe hin und wieder bewußt das „wir" benutzt, um zum Beispiel an Voraussetzungen zu erinnern („Wir haben gesehen, daß ..."), bin aber meistens beim „ich" geblieben, um klarzustellen, daß auch die mathematische Theorie auf Entscheidungen beruht, die von anderen Menschen anders hätten getroffen werden können.

Alle von mir benutzten Resultate anderer Autor/innen wurden explizit hereinzitiert. Der einzige Fall, in dem ich einen Beweis völlig analog zu einem anderen Autor geführt habe, ist deutlich gekennzeichnet (Hilfssatz 9.1).

Die Arbeit enthält einige numerische Rechnungen, sowohl in den Beweisen in Teil III, als auch in den Simulationen. Die C-Programme für alle diese Rechnungen sind bei mir in MS-DOS-Format oder als Listing verfügbar. Ebenso sind die vollständigen Simulationsergebnisse von mir erhältlich. In Teil IV wurden die wesentlichen Ergebnisse aufgelistet. Wie dort vielleicht deutlich wird, enthält die vollständige Ausgabe eine große Menge an uninteressanten Informationen; ich kann nur vor ihrer Lektüre warnen.

Ich habe die meisten englischen Fachbegriffe ins Deutsche übersetzt, war jedoch nicht ganz konsequent bei Termini, für die sich keine gute Übersetzung anbot („Maximum Likelihood"). Verbreitete englische Begriffe werden bei der Einführung der deutschen Übersetzung erwähnt oder sind zumindest im Index verzeichnet. Der Index soll insbesondere dem schnellen Auffinden von Schlüsselbegriffen und der Identifikation von Abkürzungen und englischen Fachtermini dienen. Daher enthält er sehr viele interne Querverweise, aber wenig Einträge pro Begriff.

An einigen Stellen wurde die Verteilungsfunktion Φ der Standard-Normalverteilung benötigt. Für sie wurde eine Approximation mit Fehler von höchstens $7.5 * 10^{-8}$ verwendet (Formel 26.2.17 in Zelen und Severo (1964)). Normalverteilte Pseudozufallszahlen wurden mithilfe von Φ^{-1} aus rechteckverteilten Pseudozufallszahlen erzeugt. Φ^{-1} wurde nach Odeh und Evans (1974) mit einer Genauigkeit von $1.5 * 10^{-8}$ zwischen 10^{-20} und $1 - 10^{-20}$ approximiert.

1.7 Bezeichnungen

Im Zusammenhang eines linearen Regressionsmodells bezeichnet x die unabhängige Variable (Regressor). Im Falle der Regression mit Achsenabschnitt ist $x \in \mathbb{R}^{p+1}$, wobei die $p+1$.Komponente immer gleich 1 ist. Die $p+1$.Komponente des Regressionsparameters (meist β) ist dann der Achsenabschnitt. x^- bezeichnet in diesem Fall die ersten p Komponenten von x. Das bedeutet insbesondere: $\dim(\{x_i : i \in I\}) < p+1 \Leftrightarrow x_i^-, i \in I$ liegen auf einer gemeinsamen $p-1$-dimensionalen Hyperebene

$$H := \{x^- \in \mathbb{R}^p : \alpha' x^- = a\}, \quad \mathbb{R}^p \ni \alpha \neq 0,$$

wobei $\langle A \rangle$ die lineare Hülle von A bezeichnet. $\mathcal{H}_p$ ist die Menge der $p-1$-dimensionalen Hyperebenen des $\mathbb{R}^p$. Im Falle der Regression ohne Achsenabschnitt ist $x \in \mathbb{R}^p$.

$y \in \mathbb{R}$ bezeichnet die abhängige Variable. Je nach Zusammenhang können x und y Zufallsgrößen, deren Realisation oder Unbekannte in einer Funktionsdefinition sein. Nur im Falle der Verwechslungsgefahr bezeichne ich die zugehörigen Zufallsvariablen mit X, Y. Diese Notation scheint mir übersichtlicher zu sein als die Verwendung unterschiedlicher Bezeichner. Die Gefahr von Mißverständnissen sollte klein sein. Für eine Zufallsvariable u bezeichnet $\mathcal{L}(u)$ die Verteilung von u. Für die n-fache (unabhängige) Durchführung eines Zufallsexperimentes sei $z_i := (x_i', y_i)', \mathbf{Z} := (z_1, \ldots, z_n)', x_i = (x_{i1}, \ldots, x_{ip}, 1)' \in \mathbb{R}^p \times \{1\}, y_i \in \mathbb{R}$ für $i = 1, \ldots, n$. In diesem Fall tauchen häufig Indikatorvektoren $g = (g_1, \ldots g_n) \in \{0,1\}^n$ auf. Dann ist $n(g) := \sum_{i=1}^n g_i$ und $\mathbf{Z}(g) := \left(z_{j_1}', \ldots, z_{j_{n(g)}}'\right)'$, wobei $g_{j_i} = 1$ und die j_i paarweise verschieden seien für $i = 1, \ldots, n(g)$. $y(g)$ und $\mathbf{X}(g)$ seien entsprechend definiert. Für $(x', y)'$ wird meistens (x, y) geschrieben, entsprechend $(\beta', \sigma^2)'$, wobei σ^2 im Falle normalverteilter Störterme meistens die Störvarianz bezeichnet.

Allgemein werden Matritzen durch fettgedruckte Großbuchstaben bezeichnet. $\mathbf{I}_d$ sei die d-dimensionale Einheitsmatrix. Nullvektoren beliebiger Dimension werden mit 0 bezeichnet.

Verteilungen werden, abhängig vom Argument, mit demselben Buchstaben bezeichnet wie ihre zugehörigen Verteilungsfunktionen. Ausnahme: $\Phi_{(a,S)}$ ist die Verteilungsfunktion der Normalverteilung $\mathcal{N}_{(a,S)}$ mit Mittelwertvektor a und Kovarianzmatrix S, $\varphi_{(a,S)}$ ist die entsprechende Dichte. Wird der untere Index weggelassen, handelt es sich um

die $\mathcal{N}_{(0,1)}$-Verteilung. δ_a sei das Dirac-Maß in a. $\mathcal{P}_d$ sei die Menge aller Verteilungen auf $(\mathbb{R}^d, \mathcal{B}^d)$, wobei $\mathcal{B}$ die Borel-σ-Algebra ist. Ist für die Regressoren $\mathcal{L}(x) = G \in \mathcal{P}_{p+1}$, wobei $\mathcal{L}(x_{p+1}) = \delta_1$, so wird häufig $x_{p+1} = 1$ als nichtstochastisch betrachtet und G als Verteilung aus $\mathcal{P}_p$ geschrieben, und entsprechend $\mathcal{L}(x, y) \in \mathcal{P}_{p+1}$ trotz $(x, y) \in \mathbb{R}^p \times \{1\} \times \mathbb{R}$. $\mathcal{J}(T)$ sei die Menge aller Verteilungen mit endlichem Träger auf einer Menge T, $S(J)$ sei der Träger von J für $J \in \mathcal{J}(T)$. G bezeichnet im Falle stochastischer Regressoren normalerweise die Regressorenverteilung.

"$t \leq u$" bedeutet komponentenweise $\leq$, falls t, u mehrdimensionale Vektoren sind. Abbildungen werden vollständig mit $A : U \mapsto B, u \mapsto b$ oder abkürzend mit $A(\bullet)$ notiert, wobei U der Urbildraum und B der Bildraum ist. Indikatorfunktionen werden als $1(\text{Aussage})$ notiert. „$A_n \searrow A$" bedeutet, daß (A_n) mit $n \to \infty$ monoton fällt (bei Mengen: absteigt) und gegen A konvergiert. Ich unterscheide $\mathbb{R}_0^+ := [0, \infty)$ und $\mathbb{R}^+ := (0, \infty)$.

Ist von der „Anzahl der Parameter" die Rede, so ist die Anzahl der rellen Komponenten der Parameter gemeint, also zum Beispiel $p + 2$ für $(\beta, \sigma^2) \in \mathbb{R}^{p+2}$. Das „Dach" $(\hat{a})$ für einen Modellparameter a bezeichnet außer in den Abschnitten über Identifizierbarkeit einen Schätzer des Parameters.

$E_P(f(x))$ bezeichnet den Erwartungswert von $f(x)$ wenn $\mathcal{L}(x) = P$. Falls Verwechslungsgefahr ausgeschlossen ist, schreibe ich manchmal auch $E(x)$ oder $E(P)$ für $E_P(x)$. Bei der Varianz „Var" wird analog vorgegangen. Zu einer gegebenen Verteilung P ist $P^n := \bigotimes_{i=1}^n P$.

Auf Seite 171 findet sich ein Symbolverzeichnis.

Teil I
Mischungen linearer Regressionen

2 Modellierung

In diesem Abschnitt führe ich stochastische Modelle für Cluster linearer Regressionsdaten ein. Diese Modelle sind eine Art Idealfall für die in dieser Arbeit behandelten Verfahren: Sie modellieren den Fall, daß jeder Datenpunkt von einer linearen Regressionsverteilung aus einer endlichen Mischung erzeugt wurde.

Man kann ein Experiment, das Cluster linearer Regression erzeugen soll, auf unterschiedliche Weise modellieren. Es sollen Verteilungen für $y \in (\mathbb{R}, \mathcal{B})$ bzw. $(x, y) \in \left(\mathbb{R}^{p+1} \times \mathbb{R}, \mathcal{B}^{p+2}\right)$ bei stochastischen Regressoren x gemischt werden, die folgende Situation beschreiben:

$$\begin{aligned} y_i &= x_i'\beta + u_i, \qquad \mathcal{L}(u_i) = \mathcal{N}_{0,\sigma^2} \text{ i.i.d. für } i \in I, \\ y_i &\in \mathbb{R}, \qquad x_i = (x_{i1}, \ldots, x_{ip}, 1)' \in \mathbb{R}^p \times \{1\}, \quad \beta \in \mathbb{R}^{p+1}, \end{aligned} \tag{2.1}$$

u_i stochastisch unabhängig von x_i (sofern letzteres stochastisch ist), I Indexmenge (zum Beispiel $I = \{1, \ldots, n\}$). Dabei bezeichnet die $p+1$. Komponente von β den Achsenabschnitt. x_i heißt „Regressor", y_i heißt „abhängige Variable", u_i heißt „Störterm", $\mathrm{Var}(u_i)$ Störvarianz.

Die Normalverteilungsvoraussetzung für den Störterm u_i wird für Teile der Identifizierbarkeitstheorie in Abschnitt 6 benötigt. In Teil III wird manchmal eine andere Voraussetzung für die Verteilung des Störterms verwendet. Er soll aber auf jeden Fall symmetrisch um 0 verteilt sein.

Es werden Modelle mit stochastischen und festen Regressoren unterschieden. Weiterhin gibt es zwei gängige Arten von Verteilungen, die Cluster erzeugen: Mischmodelle und Modelle mit fester Zuordnung („Fixed Partition"; eine Übersicht über stochastische Methoden der Clusteranalyse findet man zum Beispiel in Bock (1996)).

Im einfachsten Modell sind die Regressoren fest vorgegeben und die Verteilung der abhängigen Variable ist eine Mischung univariater Normalverteilungen:

Modell 1 (Feste Regressoren, Mischmodell)

$$\mathcal{L}\left((y_i)_{i\in I}\right) = \bigotimes_{i\in I} F_{x_i,J}, \text{ wobei}$$

$$F_{x,J}(y) = \int_{T_f} \Phi_{0,\sigma^2}(y - x'\beta) dJ(\beta, \sigma^2), \quad T_f := \mathbb{R}^{p+1} \times \mathbb{R}_0^+;$$

I Indexmenge, x_i, y_i wie in (2.1), $J \in \mathcal{J}(T_f)$.

Bemerkung 2.1 *Ich habe mich hier für die Schreibweise von Mischverteilungen als Integrale bezüglich einer diskreten Verteilung J auf dem Parameterraum entschieden. Das macht die Formulierungen in den Abschnitten über Identifizierbarkeit kürzer. Die häufiger verwendete Schreibweise (zum Beispiel in Titterington, Smith und Makov (1985))*

macht dagegen deutlicher, wie die interessierenden Parameter aussehen:

$$F_{x,J}(y) = \sum_{j=1}^{s} \epsilon_j \Phi_{0,\sigma_j^2}(y - x'\beta_j)$$

$$s := |S(J)|, \quad \epsilon_j := J\{(\beta_j, \sigma_j^2)\}, \text{ also}$$

$$\sum_{j=1}^{s} \epsilon_j = 1, \quad \epsilon_j > 0, \quad (\beta_j, \sigma_j^2) \text{ paarweise verschieden für } j = 1, \ldots, s.$$

Diese Schreibweise läßt sich auch auf Modell 3 übertragen.

Die übliche Interpretation für feste Regressoren ist, daß die Designer/in des Experimentes die Werte der x_i selbst gewählt hat (zum Beispiel Einflußgrößen in der Physik). Man kann die Indexmenge I dann auf unterschiedliche Weise interpretieren:

- Falls $I = \{1, \ldots, n\}$, kann n als Stichprobenumfang bei der Modellierung eines konkreten Experimentes interpretiert werden. In diesem Fall läge eine Realisation von $(y_i)_{i \in I}$ vor.
- Da ein Modell aber angeben soll, wie die Daten entstehen, und nicht, welche Daten konkret beobachtet wurden, kann n auch die Anzahl der (eventuell paarweise verschiedenen) x_i sein, die zum Design gehören und für die mehrere Beobachtungen erhoben werden können. Man könnte sich also auch mehrere unabhängige Realisationen von $(y_i)_{i \in I}$ vorstellen.
- Die x_i, $i \in I$, können auch die Regressoren sein, die theoretisch hätten gewählt werden können, wie bei der Modellierung eines stochastischen Prozesses oder einer Zeitreihe. In diesem Fall wäre auch $I = \mathbb{N}$ oder gar $x_i = i$, $I = \mathbb{R}^{p+1}$ sinnvoll, wobei man zur Parameterschätzung dann immer nur eine Auswahl der (x_i, y_i) zur Verfügung hätte.

Das Modell 1 modelliert die y als generiert von der Verteilung $\mathcal{N}_{(x'\beta,\sigma^2)}$ mit Wahrscheinlichkeit $J\{(\beta, \sigma^2)\}$. Das bedeutet: Die Wahrscheinlichkeit dafür, daß y_i aus einer Mischungskomponente mit einem bestimmten Parameter $(\beta_0, \sigma_0^2) \in S(J)$ stammt, ist unabhängig von i und x_i.

Bemerkung 2.2 *Die Unabhängigkeit der Zugehörigkeit der Datenpunkte (x_i, y_i) zu einer Mischungskomponente bzw. einem Cluster vom Regressor x_i wird in dieser Arbeit im folgenden als „Zuordnungsunabhängigkeit" bezeichnet. Da die Regressoren hier nicht immer stochastisch sind, ist die Zuordnungsunabhängigkeit im allgemeinen nicht als stochastische Unabhängigkeit formalisierbar.*

Das ist eine Einschränkung, die in vielen Anwendungssituationen nicht sinnvoll ist. Zum Beispiel wird in der Literatur die Situation, daß die y_i abhängig von der Zeit oder den x_i aus unterschiedlichen Verteilungen kommen, als „Changepoint-Regression" behandelt (siehe Abschnitt 3.1). Auch für den Telefondatensatz gibt es offenbar einen solchen Zusammenhang. Allgemeine Situationen, in denen irgendeine Abhängigkeit zwischen i und der Verteilung von y_i besteht, werden durch das Fixed Partition Model erfaßt. Dabei wird die Mischverteilung J durch unbekannte Parameter $\gamma(i)$, $i \in I$, ersetzt, die für

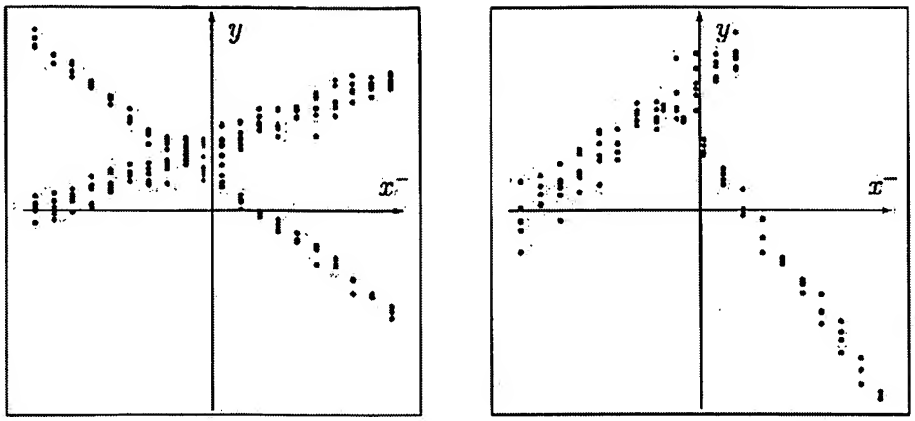


Abbildung 3: Clusterzugehörigkeit unabhängig / abhängig von x

jedes (x_i, y_i) die Verteilung angeben, die den Punkt generiert hat:

Modell 2 (Feste Regressoren, Fixed Partition Model)

$$\mathcal{L}((y_i)_{i\in I}) = \bigotimes_{i\in I} F_{x_i,\gamma(i)}, \text{ wobei}$$
$$\gamma : I \mapsto \mathbb{R}^{p+1} \times \mathbb{R}_0^+, \quad |\gamma(I)| < \infty,$$
$$F_{x,(\beta,\sigma^2)}(y) = \Phi_{0,\sigma^2}(y - x'\beta), \quad (\beta,\sigma^2) \in \gamma(I),$$

I Indexmenge, x_i, y_i wie in (2.1). Für alle $i \in I$ sei $\beta(i)$ die Projektion von $\gamma(i)$ auf die ersten $p+1$ und $\sigma^2(i)$ die Projektion auf die letzte Komponente.

Die Annahme fester Regressoren kann in einigen Anwendungen unrealistisch sein. Zum Beispiel interessiert man sich in der Ökonomie oder Psychologie für die Relationen zwischen (mehr oder weniger genau meßbaren) Eigenschaften von Individuen, die zufällig ausgewählt werden. In diesem Fall kann eine einzelne Beobachtung durch ein reines Mischmodell modelliert werden, aus dem die Stichprobe dann unabhängig identisch gezogen wird. Man umgeht also die Modellierung der Indexmenge I:

Modell 3 (Stochastische Regressoren, Mischmodell) *Es seien* $(x_i, y_i)_{i\in I}$ *unabhängig identisch verteilt mit* $\mathcal{L}(x_i, y_i) = F_J$,

$$F_J(x,y) = \int_{T_s} F(x,y,\theta)dJ(\theta), \text{ wobei}$$
$$F(x,y,\theta) = \int 1(t \le x)\Phi_{0,\sigma^2}(y - t'\beta)dG(t),$$
$$\theta := (\beta, \sigma^2, G) \in T_s := I\!R^{p+1} \times I\!R_0^+ \times \mathcal{G}, \quad J \in \mathcal{J}(T_s),$$

$\mathcal{G} \subset \mathcal{P}_{p+1}$ *sei meßbar (siehe die folgende Bemerkung), wobei* $\mathcal{L}(x_{p+1}) = \delta_1 \ \forall G \in \mathcal{G}$. *Aus Gründen der Einfachheit wird* G *im folgenden meist als p-dimensionale Verteilung behandelt und* $x_{p+1} = 1$ *sei fest.*

Bemerkung 2.3 *Die Verteilung* G *der Regressoren ist hier ein Parameter des Modells. Um klarzustellen, daß man dadurch keine Meßbarkeitsprobleme bekommt, zitiere ich kurz einige Ergebnisse aus Abschnitt 12 von Hinderer (1970):*
Man definiert eine σ*–Algebra* $\mathcal{B}$ *auf* $\mathcal{P}_p$ *wie folgt:*

$$\mathcal{B} := \sigma(\{\{P \in \mathcal{P}_p : P(A) \in B\} : A \in I\!B^p, B \in I\!B\}).$$

Dann gilt:

- $(\mathcal{P}_p, \mathcal{B})$ *ist ein Standard Borel-Raum, d.h.* $\mathcal{B}$ *wird von einer Topologie erzeugt, bezüglich derer* $\mathcal{P}_p$ *ein polnischer Raum ist (Satz 12.13). Daraus folgt:*
- $\mathcal{B}$ *enthält alle einelementigen Teilmengen aus* $\mathcal{P}_p$ *(S. 87) und damit die Projektionen auf* $\mathcal{P}_p$ *der Träger der Elemente aus* $\mathcal{J}(T_s)$.
- *Sei* $u : I\!R^{q+p} \mapsto I\!R$ *eine meßbare Abbildung, die von oben oder unten beschränkt ist. Dann ist die Abbildung*

$$t : (x, P) \mapsto \int u(\bullet, x)dP, \quad (x, P) \in I\!R^q \times \mathcal{P}_p$$

$I\!B^q \otimes \mathcal{B}$-meßbar (Lemma 12.2). Die Voraussetzungen an u *werden von Verteilungsfunktionen erfüllt.*

Also ist $\mathcal{J}(T_s)$ *wohldefiniert,* $\int F(x,y,\bullet)dJ$ *wohldefiniert auf* $(T_s, I\!B^{p+1} \otimes \mathcal{B})$ *und meßbar.*

Da jede Mischungskomponente eine eigene Regressorenverteilung enthält, ist die Zuordnung der Beobachtungen zu den Mischungskomponenten im allgemeinen abhängig von den Regressoren. Sind allerdings die G für alle $\theta \in S(J)$ gleich, kann auch Zuordnungsunabhängigkeit modelliert werden.

Für die Theorie in Teil III werde ich normalerweise das Modell mit stochastischen Regressoren verwenden, da die Produktbildung über I entfällt, was für Rechnungen mit Schätzerfunktionalen am leichtesten handhabbar ist. Für die Berechnung von konventionellen Parameterschätzern (wie in Abschnitt 3.3) ist das Modell jedoch problematisch wegen der Abhängigkeit von den normalerweise unbekannten Verteilungen G der Regressoren.

Es ist auch möglich, stochastische Regressoren mit dem Fixed Partition-Ansatz zu modellieren. Das daraus resultierende Modell ist mathematisch am wenigsten handhabbar und mir fallen keine realistischen Situationen ein, in denen die Interpretation ein solches Modell erzwingen würde. Ich werde das Modell hier trotzdem vorstellen, denn aus verschiedenen Gründen (siehe Abschnitt 14.2) habe ich es zur Datengenerierung in den Simulationen verwendet. In den Abschnitten über Parameterschätzung und Identifizierbarkeit werde ich aber nicht darauf eingehen.

Modell 4 (Stochastische Regressoren, Fixed Partition Model)

$$\mathcal{L}\left((x_i, y_i)_{i\in I}\right) = \bigotimes_{i\in I} F_{\gamma(i)}, \text{ wobei}$$

$$\gamma : I \mapsto \mathbb{R}^{p+1} \times \mathbb{R}_0^+ \times \mathcal{G}, \quad |\gamma(I)| < \infty,$$

$$F_{(\beta,\sigma^2,G)}(x, y) = \int 1(t \le x)\Phi_{0,\sigma^2}(y - t'\beta)dG(t), \quad (\beta, \sigma^2, G) \in \gamma(I),$$

I Indexmenge, x_i, y_i wie in (2.1), $\mathcal{G} \subset \mathcal{P}_{p+1}$ meßbar, wobei $\mathcal{L}(x_{p+1}) = \delta_1 \; \forall G \in \mathcal{G}$.

Bemerkung 2.4 *Das lineare Regressionsproblem ist äquivariant unter linearen Transformationen der Form*

$$D \in \mathcal{D} := \left\{D:\ \mathbb{R}^{p+2} \mapsto \mathbb{R}^{p+2},\ (x, y) \mapsto ((\Gamma x), (ay + x'b))\right\}, \tag{2.2}$$

wobei $\Gamma \in \mathbb{R}^{(p+1)^2}$ invertierbar, $a \in \mathbb{R} \setminus \{0\}$, $b \in \mathbb{R}^{p+1}$. An $\mathcal{D}$ kann weiterhin vorausgesetzt werden, daß $(0, \ldots, 0, 1)$ die letzte Zeile von Γ ist, um sicherzustellen, daß $(\Gamma x)_{p+1} = 1$. Auch in diesem Fall ist zu $D \in \mathcal{D}$ immer $D^{-1} \in \mathcal{D}$, da die letzte Zeile der Inversen von Γ dann auch $(0, \ldots, 0, 1)$ ist.

Das heißt, daß ein Regressionsmodell der Form (2.1) bei Anwendung von D auf (x, y) in ein Modell derselben Form übergeht, bei dem β sich in $(\Gamma^{-1})'(a\beta + b)$ und σ^2 sich in $a^2\sigma^2$ verwandelt.

In der Schreibweise aus Modell 2:

$$\begin{aligned} &F_{\Gamma x,((\Gamma^{-1})'(a\beta+b),a^2\sigma^2)}(ay + x'b) = \\ &= \Phi_{0,a^2\sigma^2}\left(ay + x'b - (\Gamma x)'(\Gamma^{-1})'(a\beta + b)\right) = \\ &= \Phi_{0,a^2\sigma^2}\left(a(y - x'\beta)\right) = \Phi_{0,\sigma^2}(y - x'\beta) = F_{x,(\beta,\sigma^2)}(y). \end{aligned}$$

Durch Einsetzen dieser Gleichung folgt auch die Äquivarianz bzgl. D für die vier in diesem Abschnitt eingeführten Modelle:

Modell 1: *Ist $(y_i)_{i\in I}$ verteilt wie in Modell 1, so gilt für Γ, a, b wie oben:*

$$\mathcal{L}\left((ay_i + x_i'b)_{i\in I}\right) = \bigotimes_{i\in I} F_{\Gamma x_i, J^*}, \text{ wobei}$$

$$J^*((\Gamma^{-1})'(a\beta + b), a^2\sigma^2) := J(\beta, \sigma^2)\ \forall (\beta, \sigma^2) \in T.$$

Modell 2: *Ist $(y_i)_{i\in I}$ verteilt wie in Modell 2, so gilt für Γ, a, b wie oben:*

$$\mathcal{L}\left((ay_i + x_i'b)_{i\in I}\right) = \bigotimes_{i\in I} F_{\Gamma x_i, \gamma^*(i)}, \text{ wobei}$$
$$\gamma^*(i) := ((\Gamma^{-1})'(a\beta(i) + b), a^2\sigma^2(i)) \ \forall i \in I.$$

Modell 3: *Ist $\forall G \in \mathcal{G}$ auch $G_\Gamma \in \mathcal{G}$, wobei*

$$G_\Gamma(B) := G\{x : \Gamma x \in B\} \ \forall B \in \mathbb{B}^{p+1},$$

und ist $\mathcal{L}(x, y) = F_J$ wie in Modell 3, dann gilt für Γ, a, b wie oben:

$$\mathcal{L}(\Gamma x, ay + x'b) = F_{J^*}, \text{ wobei}$$
$$J^*((\Gamma^{-1})'(a\beta + b), a^2\sigma^2, G_\Gamma) := J(\beta, \sigma^2, G) \ \forall (\beta, \sigma^2, G) \in T.$$

Modell 4: *Ist $\forall G \in \mathcal{G}$ auch $G_\Gamma \in \mathcal{G}$, und ist $(x_i, y_i)_{i\in I}$ verteilt wie in Modell 4, so gilt für Γ, a, b wie oben:*

$$\mathcal{L}\left((\Gamma x_i, ay_i + x_i'b)_{i\in I}\right) = \bigotimes_{i\in I} F_{\gamma^*(i)}, \text{ wobei}$$
$$\gamma^*(i) := ((\Gamma^{-1})'(a\beta(i) + b), a^2\sigma^2(i), G(i)_\Gamma) \ \forall i \in I.$$

Von allen später diskutierten Schätzverfahren werde ich zeigen, daß sie sich bei linearer Transformation der Daten ebenfalls äquivariant verhalten. Das bedeutet, daß es bei der Berechnung der Schätzer möglich ist, die Daten zu transformieren, zum Beispiel auf Mittelwertvektor 0 und Kovarianzmatrix $\mathbf{I}$. Theoretische Resultate und die Ergebnisse aus Simulationen für feste Parameter lassen sich durch lineare Transformation verallgemeinern.

3 Ansätze zur Analyse der Modelle

3.1 Wechselpunktprobleme

Unter Wechselpunkt- („Changepoint"-)Problemen werden Situationen verstanden, in denen ein System ab einem bestimmten Zeitpunkt oder ab einem bestimmten Wert einer beobachtbaren Einflußgröße von einem Zustand in einen anderen wechselt. Im Regressions-Zusammenhang würde sich an einem solchen Wechselpunkt der Zusammenhang zwischen x und y verändern, also die Regressions- und/oder Skalenparameter. Mit Modell 2 können Wechselpunktprobleme im Prinzip modelliert werden, wenn sich die Clusterzugehörigkeit γ abhängig von zum Beispiel i oder einer x-Komponente ändert. Allerdings unterscheiden sich Wechselpunktprobleme von der Situation, die uns in dieser Arbeit interessiert, denn man hat dort eine zusätzliche Information: Es ist bekannt, wovon die Clusterzugehörigkeit abhängt. Daher kann die Anwendung der hier diskutierten Verfahren nur dann sinnvoll sein, wenn man sich nicht sicher ist, ob man es mit einer Wechselpunkt-Situation zu tun hat. Anderenfalls würde vorhandene Information ungenutzt bleiben.

Diese Information ermöglicht bessere theoretische Resultate bei der Parameterschätzung, als in der allgemeinen Clustersituation möglich sind. Daher gibt es eine Fülle von

Literatur über Wechselpunkt-Regression. Einen Überblick über Verfahren und Resultate geben Krishnaiah und Miao (1988). Der aktuellste mir bekannte Artikel ist Huskova (1996). Der erste Artikel über verschiedene Klassen linearer Regression (Quandt (1958)) formuliert ebenfalls ein Wechselpunktproblem.

Unter bestimmten Voraussetzungen läßt sich im Wechselpunktproblem auch die Anzahl der Wechselpunkte konsistent schätzen. Yao (1988) verwendet dazu das Schwarz'sche Kriterium (auch „Bayes'sches Informationskriterium" BIC; Schwarz (1978) schlägt es ursprünglich für die Modellwahl in der Regression vor):

$$\ln L_n(s) - \frac{1}{2}k(s)\ln n \stackrel{!}{=} \max_{s\in N} \tag{3.1}$$

Dabei sei s die Anzahl der Wechselpunkte (später: Anzahl der Cluster), n die Anzahl der Beobachtungen, $L_n(s)$ das Maximum der Likelihoodfunktion im durch s definierten Modell und $k(s)$ die Anzahl der zu schätzenden Parameter.

Für die allgemeine Regressions-Clusteranalyse gibt es keine vergleichbaren Konsistenzresultate. Die Verwendung des Schwarz'schen Kriteriums führt aber auch dort zu brauchbaren Ergebnissen, wie sich später zeigen wird.

3.2 Kleinste Quadrate

Gegeben sei nun die Situation, daß keine weiteren Informationen darüber vorhanden sind, wodurch unterschiedliche Cluster linearer Regression verursacht werden. Wenn die Anzahl der Cluster s bekannt ist, kann man die Methode der kleinsten Quadrate (KQ) anwenden, um die Regressionsparameter zu schätzen[4]:

$$\sum_{i=1}^{n}\sum_{j=1}^{s} 1(\zeta(i)=j)(y_i-\beta_j' x_i)^2 \stackrel{!}{=} \min_{\zeta,\beta_1,\ldots,\beta_s} \tag{3.2}$$

Dabei sei s wieder die Anzahl der Cluster, n die Anzahl der Beobachtungen und $\zeta : \{1,\ldots,n\} \mapsto \{1,\ldots,s\}$ eine Abbildung, die die Clusterzugehörigkeit der einzelnen Punkte angibt. Für dieses Problem konvergiert folgender Algorithmus zumindest gegen ein lokales Minimum der Zielfunktion:

1. Beginne mit einer Startpartition $\hat{\zeta}$.
2. Berechne die KQ-Schätzer für die einzelnen Cluster.
3. Ordne jeden Punkt dem Cluster zu, in dem er den kleinsten Residuumsbetrag liefert.

Dieser Algorithmus wurde zuerst vorgeschlagen von Bock (1969)[5]. Spaeth (1979) weist darauf hin, daß dieser Algorithmus manchmal Cluster generiert, deren Regressoren auf einer gemeinsamen p-dimensionalen Hyperebene des $\mathbb{R}^{p+1}$ liegen, zum Beispiel Cluster mit weniger als $p+1$ Punkten, so daß der KQ-Schätzer nicht mehr berechenbar ist. Er schlägt einen Austauschalgorithmus vor. Die Arbeit von Spaeth wurde insbesondere in den Wirtschaftswissenschaften beachtet. Es gibt eine Reihe von Verallgemeinerungen

[4] Bis zum Ende von Abschnitt 3 sei $I = \{1,\ldots,n\}$.

[5] Bock beweist auch die Konvergenz, sogar allgemeiner für $y \in \mathbb{R}^t, t \geq 1$ und entsprechende x, β.

auf kompliziertere Regressionscluster-Situationen, die in der Ökonometrie auftauchen (siehe zum Beispiel Wedel und Steenkamp (1991)). Da die Zielfunktion mit steigender Clusterzahl fällt, kann man die Zahl der Cluster nicht mit dem KQ-Kriterium schätzen. Charles (1979) schlägt vor, die Anzahl der Cluster so zu wählen, daß das mit der Anzahl der Cluster steigende Verhältnis der Varianz zwischen den Clustern zur Gesamtvarianz einer globalen Regression einen „Knickpunkt" hat.

Die Berechnung von Regressionsparametern mit der KQ-Methode ignoriert in den vorgeschlagenen Modellen die normalerweise unbekannten Skalenparameter $\sigma_1^2, \ldots, \sigma_s^2$. In Abschnitt 3.4 wird sich zeigen, daß die KQ-Methode in Modell 2 im Fall $\sigma_1^2 = \ldots = \sigma_s^2$ äquivalent zur ML-Schätzung ist.

Einen alternativen KQ-Ansatz liefert Jajuga (1986). Er setzt in Modell 3 voraus, daß die Regressoren - abgesehen vom Achsenabschnitt - in allen Clustern normalverteilt sind, d.h. $\mathcal{G} = \{\mathcal{N}_{(\eta,\mathbf{AA}')} | \eta \in I\!R^p, \mathbf{A} \text{ invertierbar } p \times p\}$. In diesem Fall sind die gemeinsamen Verteilungen $F(\bullet, \beta, \sigma^2, G)$ reparametrisierte $p+1$-dimensionale Normalverteilungen (siehe Beweis von Satz 6.7). Jajuga schlägt vor, zuerst eine Clusteranalyse für das mehrdimensionale Lokationsproblem durchzuführen. Das kann zum Beispiel mit Hilfe einer ML-Schätzung im Mischmodell für $p+1$-dimensionale Normalverteilungen geschehen. Innerhalb der Cluster wird dann der normale KQ-Regressionsschätzer berechnet. Die statistischen Eigenschaften dieses Verfahrens sind unklar. Die Regressoren als normalverteilt vorauszusetzen, ist eine starke Einschränkung. Andererseits werden dadurch Identifizierbarkeitsprobleme ausgeschlossen (siehe Satz 6.7).

3.3 Parameterschätzung im Mischmodell

Wir betrachten nun Modell 1 in der Formulierung aus Bemerkung 2.1. Sei vorerst die Anzahl der Mischungskomponenten s bekannt. Es ergibt sich folgende Loglikelihoodfunktion:

$$\ln L_n[s, (\beta_1, \sigma_1^2, \epsilon_1), \ldots, (\beta_s, \sigma_s^2, \epsilon_s), \mathbf{Z}] = \\ = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{s} \epsilon_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left[\frac{-(y_i - \beta_j' x_i)^2}{2\sigma_j^2} \right] \right). \tag{3.3}$$

Den ersten Vorschlag zur Berechnung des ML-Schätzers für $((\beta_1, \sigma_1^2, \epsilon_1), \ldots, (\beta_s, \sigma_s^2, \epsilon_s))$ machte Quandt (1972). Hosmer (1974) löst das Problem für $p = 1$ und $s = 2$ mit einem Newton-Algorithmus. Mit auch für höhere Dimensionen und Clusterzahlen sinnvollem Aufwand ist $\ln L_n$ mit dem EM-Algorithmus nach Dempster, Laird und Rubin (1977) lokal zu maximieren. Dabei wird wie folgt vorgegangen:

Schritt 1: Wähle eine Startpartition $(\hat{\epsilon}_{ij})_{i=1,\ldots,n; j=1,\ldots,s}$, wobei $\sum_{j=1}^{s} \hat{\epsilon}_{ij} = 1 \quad \forall i$, $\hat{\epsilon}_{ij} \geq 0 \quad \forall i, j$.

Schritt 2:

$$\hat{\epsilon}_j = \frac{\sum_{i=1}^{n} \hat{\epsilon}_{ij}}{n} \quad \forall j = 1, \ldots, s.$$

Schritt 3: $\hat{\beta}_j$ sei der KQ-Schätzer für $j = 1, \ldots, s$ mit durch $(\hat{\epsilon}_{ij})_{i=1,\ldots,n}$ gewichteten Beobachtungen, d.h.

$$\hat{\beta}_j = (\mathbf{X}' \text{diag}(\hat{\epsilon}_{ij})_{i=1,\ldots,n} \mathbf{X})^{-1} \mathbf{X}' \text{diag}(\hat{\epsilon}_{ij})_{i=1,\ldots,n} y.$$

Schritt 4:

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_{ij}(y_i - \hat{\beta}_j' x_i)^2}{n\hat{\epsilon}_j} \quad \forall j = 1, \ldots, s.$$

Schritt 5:

$$\hat{\epsilon}_{ij} = \frac{\hat{\epsilon}_j \varphi_{(x_i'\hat{\beta}_j, \hat{\sigma}_j^2)}(y_i)}{\sum_{l=1}^s \hat{\epsilon}_l \varphi_{(x_i'\hat{\beta}_l, \hat{\sigma}_l^2)}(y_i)} \quad \forall i = 1, \ldots, n, j = 1, \ldots, s$$

Schritt 6: Abbruch, wenn Vergrößerung von $\ln \hat{L}_n$ im letzten Durchgang kleiner als eine vorgegebene Schranke, sonst weiter bei Schritt 2.

Diesen Ansatz leiten DeSarbo und Cron (1988) her. $\hat{\epsilon}_{ij}$ kann als die geschätzte Wahrscheinlichkeit, daß Punkt i von Mischungskomponente j erzeugt ist, interpretiert werden. Diese Interpretation kann angewendet werden, um die Punkte auch in die Mischungskomponenten zu klassifizieren (siehe (15.1) im Abschnitt 15.1.2).

Bemerkung 3.1 *Die ML-Methode hat folgendes Problem:* $\ln \hat{L}_n$ *ist unbeschränkt, falls* $\hat{\sigma}_j^2 \to 0$ *für eine Komponente* j*:*
Sei zum Beispiel $s = 2$*. Man wähle* $\hat{\beta}_1$ *so, daß* $y_1 - \hat{\beta}_1' x_1 = 0$*.* $\hat{\beta}_2, \hat{\sigma}_2^2 \in I\!R^{p+1} \times I\!R^+$ *seien beliebig und fest. Dann ist für* $\hat{\epsilon}_{11} = 1, \hat{\epsilon}_{i1} = 0$ *für* $i = 2, \ldots, n$*,* $\hat{\epsilon}_1 = \frac{1}{n}$*,* $\hat{\epsilon}_{i2} = 1 - \hat{\epsilon}_{i1}$ *für* $i = 1, \ldots, n$*,* $\hat{\epsilon}_2 = \frac{n-1}{n}$*:*

$$\ln L_n[2, (\hat{\beta}_1, \hat{\sigma}_1^2, \hat{\epsilon}_1), (\hat{\beta}_2, \hat{\sigma}_2^2, \hat{\epsilon}_2), \mathbf{Z}] =$$
$$= \ln\left(\frac{1}{n\sqrt{2\pi\hat{\sigma}_1^2}} \exp(0)\right) + \sum_{i=2}^n \ln\left(\frac{n-1}{n\sqrt{2\pi\hat{\sigma}_2^2}} \exp\left[\frac{-(y_i - \hat{\beta}_2' x_i)^2}{2\hat{\sigma}_2^2}\right]\right).$$

Der erste Summand konvergiert gegen ∞ *für* $\hat{\sigma}_1^2 \to 0$*.*

Für die Theorie bedeutet das, daß man nicht an einer globalen Maximalstelle der Likelihood interessiert ist, sondern nur an einem lokalen Maximum, das eine gewisse Entfernung vom durch $\hat{\sigma}^2 = 0$ gegebenen Rand des Parameterraums hat. Für die Praxis schlagen DeSarbo und Cron (1988) vor, in Schritt 4 zum Beispiel $\hat{\sigma}_j^2 := 0.1$ oder eine andere vorgegebene untere Schranke zu setzen, falls nach der regulären Berechnung $\hat{\sigma}_j^2$ unter dieser Schranke liegt. Weiterhin kann es passieren, daß weniger als $p+1$ der Werte $\hat{\epsilon}_{ij}$, $i = 1, \ldots, n$ für ein j nennenswert größer als 0 sind, d.h. präziser: $\dim\langle x_i : \hat{\epsilon}_{ji} > \kappa > 0\rangle < p+1$ für sehr kleines κ, so daß es numerische Schwierigkeiten bei der Berechnung des gewichteten KQ-Schätzers $\hat{\beta}_j$ gibt. Dieser Fall muß bei der Implementierung des Algorithmus beachtet werden.

Da die Loglikelihoodfunktion unbeschränkt ist, gibt es keine einfachen asymptotischen Resultate für den ML-Schätzer. Im Lokations-Mischmodell für Normalverteilungen, wo dasselbe Unbeschränktheitsproblem auftaucht, kann aber bewiesen werden, daß eine Folge von lokalen Maxima der Loglikelihoodfunktion existiert, die konsistent und sogar asymptotisch normal ist. Titterington, Smith und Makov (1985) geben auf Seite 92 einen Überblick über die Literatur zu diesem und ähnlichen Ergebnissen. DeSarbo und Cron (1988) und Kiefer (1978) behaupten Entsprechendes auch für das lineare

Regressions-Mischmodell. Ihre Argumentationen kranken aber daran, daß entsprechende Ergebnisse für einfache Normalverteilungen zu leichtfertig auf den Regressionsfall übertragen werden. Die Tatsache, daß man es in Modell 1 mit einer Produktverteilung zu tun hat, wird ignoriert[6]. Die Parameter des Modells 1 sind unter den Voraussetzungen, die in den beiden oben erwähnten Arbeiten gemacht werden, im allgemeinen nicht einmal identifizierbar (siehe Beispiel 5.5). Das deutet darauf hin, daß man für eine korrekte Asymptotik schärfere Bedingungen bräuchte.

Quandt und Ramsey (1978) schlagen vor, die Parameter nicht über den ML-Ansatz, sondern über die momentgenerierende Funktion (MGF) zu schätzen. Dieser Ansatz ist eine Verallgemeinerung der Momentenmethode („method of moments"), die unter anderem für die Schätzung der Parameter von Lokationsmischungen verwendet wird (zum Beispiel von Day (1969)). Dabei werden so viele Stichprobenmomente berechnet, wie Parameter zu schätzen sind. Die Parameterschätzer werden dann daraus zurückgerechnet. Weil die Varianzen höherer Stichprobenmomente sehr hoch sind, schlagen Quandt und Ramsey vor, stattdessen $\nu_1, \ldots, \nu_k \in \mathbb{R}$ vorzugeben, wobei k die Anzahl der zu schätzenden Parameter ist. Dann wird

$$\sum_{j=1}^{k} \left(\sum_{i=1}^{n} (e^{\nu_j y_i} - m(x_i, \nu_j, \hat{\theta})) \right)^2$$

minimiert. Dabei ist $\hat{\theta} := ((\hat{\beta}_1, \hat{\sigma}_1^2, \hat{\epsilon}_1), (\hat{\beta}_2, \hat{\sigma}_2^2))$[7] und

$$m(x, \nu, \hat{\theta}) := \hat{\epsilon}_1 \exp\left(\hat{\beta}_1' x\nu + \nu^2 \frac{\hat{\sigma}_1^2}{2}\right) + (1 - \hat{\epsilon}_1) \exp\left(\hat{\beta}_2' x\nu + \nu^2 \frac{\hat{\sigma}_2^2}{2}\right)$$

die MGF der Verteilung von y unter gegebenem x. In der Arbeit wird Konsistenz und asymptotische Normalität für die Schätzer behauptet. Auch Quandt und Ramsey verwenden dafür eine nicht genau ausgeführte Verallgemeinerung vom Fall einer einfachen Lokationsmischung zweier Normalverteilungen. Immerhin scheinen ihre Bedingungen an die Folge $(x_i)_{i \in I}$ im Fall $s = 2$ Identifizierbarkeitsprobleme auszuschließen, auch wenn die Autoren auf diese Frage nicht explizit eingehen.

Weiterhin geben die Autoren einen Algorithmus zur Berechnung der Schätzer nach der MGF-Methode an, der offenbar recht aufwendig ist und nicht immer konvergiert. Aus letzterem Grund habe ich ihren Schätzer in meinen Simulationen nicht berücksichtigt.

Für den Regressionsfall mit $s = 2, p = 1$ ohne Achsenabschnitt schlagen Huang und Pao (1991) einen Minimum-Distanz-Schätzer vor. In diesem einfachen Fall tauchen keine Identifizierbarkeitsprobleme auf. Die Autoren leiten Konsistenz und asymptotische Normalität her, geben aber keinen Algorithmus zur Berechnung ihres Schätzers an.

Der einzige konkrete Vorschlag zur Schätzung der Anzahl der Mischungskomponenten stammt von DeSarbo und Cron (1988). Sie verwenden (ohne weitere Begründung) Akaikes Informationskriterium (AIC; siehe Akaike (1974)):

$$\ln L_n(s) - k(s) \stackrel{!}{=} \max_{s \in \mathbb{N}}, \tag{3.4}$$

Dabei seien wieder s die Anzahl der Cluster, n die Anzahl der Beobachtungen, $\ln L_n(s)$ sei die Loglikelihoodfunktion im durch s definierten Modell an der Stelle des ML-Schätzers,

[6] Kiefer (1978) spricht dieses Problem in einer Fußnote an.

[7] Quandt und Ramsey beschränken sich auf den Fall $s = 2$.

und $k(s) = (p+3)s - 1$ sei die Anzahl der zu schätzenden Parameter: für jeden Cluster $\beta \in \mathbb{R}^{p+1}, \sigma^2$ und ϵ, wobei $\epsilon_s = 1 - \sum_{i=1}^{s-1} \epsilon_i$.

In den Simulationen in Teil IV wird sich zeigen, daß das BIC (aus (3.1)) bessere Ergebnisse liefert. Zur theoretischen Rechtfertigung der beiden Kriterien gibt es nur Ergebnisse im Wechselpunktproblem (siehe Abschnitt 3.1) und für Lokations-Mischverteilungen. Dort zeigt Leroux (1992), daß sich mit Hilfe des AIC und auch des BIC die vermischende Verteilung (entsprechend J in Abschnitt 2) konsistent schätzen läßt. Damit ist impliziert, daß die Anzahl der Mischungskomponenten nicht unterschätzt wird. Allerdings impliziert konsistente Schätzung der vermischenden Verteilung nicht die konsistente Schätzung der Anzahl der Mischungskomponenten, da sich in einer beliebig kleinen Umgebung einer Mischverteilung Mischverteilungen mit einer beliebigen größeren Anzahl von Mischungskomponenten befinden.

Bemerkung 3.2 *Angenommen, ein Ergebnis wie das oben zitierte von Leroux (1992) gelte auch im Regressionsclusterfall. Dann wäre das AIC asymptotisch auf keinen Fall besser als das BIC, denn die mit dem BIC geschätzte Clusterzahl ist für $n > e^2$ immer kleiner oder gleich der mit dem AIC geschätzten Clusterzahl.*

Beweis: Sei $s_A = \arg\max_{s \in \mathbb{N}} (\ln L_n(s) - k(s))$, $n > e^2$, also $\frac{1}{2}\ln n > 1$. Dann gilt für $\mathbb{N} \ni t > s_A$ (weil nach Definition $k(t) > k(s_A)$):

$$\ln L_n(t) - \ln L_n(s_A) \leq k(t) - k(s_A) < \tfrac{1}{2}\ln n(k(t) - k(s_A)) \Rightarrow$$
$$\Rightarrow \ln L_n(t) - \tfrac{1}{2}\ln nk(t) < \max_{s \in \mathbb{N}} \left(\ln L_n(s) - \frac{1}{2}\ln nk(s)\right) \Rightarrow$$
$$\Rightarrow s_A \geq \arg\max_{s \in \mathbb{N}} \left(\ln L_n(s) - \tfrac{1}{2}\ln nk(s)\right).$$

Bemerkung 3.3 *Sei $\mathbf{Z}^D := (D(z_1), \ldots, D(z_n))$ mit D gemäß (2.2). Dann gilt für beliebige $((\beta_1, \sigma_1^2, \epsilon_1), \ldots, (\beta_s, \sigma_s^2, \epsilon_s))$:*

$$\ln L_n[s, (\beta_1, \sigma_1^2, \epsilon_1), \ldots, (\beta_s, \sigma_s^2, \epsilon_s), \mathbf{Z}] - n\ln a =$$
$$= \ln L_n[s, ((\Gamma^{-1})'(a\beta_1 + b), a^2\sigma_1^2, \epsilon_1), \ldots, ((\Gamma^{-1})'(a\beta_s + b), a^2\sigma_s^2, \epsilon_s), \mathbf{Z}^D],$$
$$\text{denn } \tfrac{1}{a}\varphi_{(x'\beta,\sigma^2)}(y) = \varphi_{((\Gamma x)'(\Gamma^{-1})'(a\beta+b),a^2\sigma^2)}(ay + x'b) \qquad (3.5)$$

für beliebige $(\beta, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$, $(x', y)' \in \mathbb{R}^{p+2}$.

Damit ist der ML-Schätzer im Mischmodell linear äquivariant. Das heißt: Ist

$$((\hat\beta_1, \hat\sigma_1^2, \hat\epsilon_1), \ldots, (\hat\beta_s, \hat\sigma_s^2, \hat\epsilon_s))$$

(lokale) Minimalstelle von $\ln L_n(s, \bullet, \mathbf{Z})$, so ist

$$[(\Gamma^{-1})'(a\hat\beta_1 + b), a^2\hat\sigma_1^2, \hat\epsilon_1), \ldots, ((\Gamma^{-1})'(a\hat\beta_s + b), a^2\hat\sigma_s^2, \hat\epsilon_s)]$$

(lokale) Minimalstelle von $\ln L_n(s, \bullet, \mathbf{Z}^D)$.

Weiterhin gilt

$$\arg\max_s[\min \ln L_n(s, \bullet, \mathbf{Z}) - k(s)] = \arg\max_s[\min \ln L_n(s, \bullet, \mathbf{Z}^D) - k(s)] \text{ und}$$
$$\arg\max_s[\min \ln L_n(s, \bullet, \mathbf{Z}) - \tfrac{1}{2}k(s)\ln n] = \arg\max_s[\min \ln L_n(s, \bullet, \mathbf{Z}^D) - \tfrac{1}{2}k(s)\ln n]$$

mit $k(s)$ gemäß (3.4), d.h. die Schätzung der Anzahl der Cluster mit dem AIC und mit dem BIC ist invariant unter Transformationen der Form (2.2).

3.4 Parameterschätzung im Fixed Partition Model

Auch im Fixed Partition Model (Modell 2) läßt sich bei bekannter Anzahl s der Cluster ein ML-Schätzer herleiten. Hier sieht die Loglikelihoodfunktion mit den Bezeichnungen aus Modell 2 folgendermaßen aus:

$$\ln L_n(s,\gamma,\mathbf{Z}) = -\tfrac{1}{2}\sum_{i=1}^{n}\left(\ln 2\pi + \ln\sigma^2(i) + \frac{(y_i-\beta(i)'x_i)^2}{\sigma(i)^2}\right) = \tag{3.6}$$

$$= -\tfrac{1}{2}\sum_{j=1}^{s}\sum_{\gamma(i)=(\beta_j,\sigma_j^2)}\left(\ln 2\pi + \ln\sigma_j^2 + \frac{(y_i-\beta_j'x_i)^2}{\sigma_j^2}\right) = \tag{3.7}$$

$$= -\frac{1}{2}\left(n\ln 2\pi + \sum_{j=1}^{s} n_j \ln\sigma_j^2\right) - \frac{1}{2}\sum_{j=1}^{s} r(j), \text{ wobei}$$

$$r(j) := \sum_{\gamma(i)=(\beta_j,\sigma_j^2)} \frac{(y_i-\beta_j'x_i)^2}{\sigma_j^2}, \quad j=1,\ldots,s, \tag{3.8}$$

$$\gamma(I) = \{(\beta_j,\sigma_j^2)|j=1,\ldots,s\},\ n_j := |\{i:\gamma(i)=(\beta_j,\sigma_j^2)\}|, j=1,\ldots,s.$$

Die (β_j,σ_j^2), $j=1,\ldots,s$ seien als paarweise verschieden vorausgesetzt, also $s=|\gamma(I)|$. Zu schätzen sind hier die $\gamma(i)$, $i=1,\ldots,n$. Das beinhaltet die Schätzung von (β_j,σ_j^2), $j=1,\ldots,s$. Über diesen Ansatz gibt es im Regressionsfall keine Literatur. Scott und Symons (1971) schätzen mit dem ML-Ansatz Partitionen von Lokationsdaten aus mehrdimensionalen Normalverteilungen.

Es ist ein Vorteil des Fixed Partition Models, daß die Cluster, die die einzelnen Daten generiert haben, zu den geschätzten Parametern gehören, so daß die Clusterzugehörigkeit der Daten mitgeschätzt wird. Im Mischmodell 1 kann die Zuordnung der Daten zu den Komponenten jedoch über die Größen $\hat{\varepsilon}_{ij}$ (siehe Abschnitt 3.3) erfolgen.

Im Regressionsfall hat das Fixed Partition Model aber noch einen weiteren Vorteil. In Modell 3 würden die ML-Schätzer von den unbekannten Regressorenverteilungen abhängen. Das Fixed Partition Model ermöglicht dagegen die Schätzung der Parameter bei Clusterzugehörigkeiten, die vom Regressor oder von i abhängig sind (siehe die Diskussion in Abschnitt 2).

Zur Berechnung des ML-Schätzers: Sei $\zeta(i) := j$, falls $\gamma(i) = (\beta_j,\sigma_j^2)$, d.h. $\zeta(i)$ gibt die Nummer des Clusters an, dem (x_i,y_i) angehört. Ich diskutiere die Schätzung der $\zeta(i)$, (β_j,σ_j^2), die gleichbedeutend zur Schätzung der $\gamma(i), i=1,\ldots,n$ ist. Aus (3.6) folgt: Für gegebene $(\hat{\beta}_j,\hat{\sigma}_j^2)$, $j=1,\ldots,s$, wird $\ln\hat{L}_n$ maximiert durch

$$\hat{\zeta}(i) = \arg\min_j \left(\ln\hat{\sigma}_j^2 + \frac{(y_i-\hat{\beta}_j'x_i)^2}{\hat{\sigma}_j^2}\right). \tag{3.9}$$

Weiter ist (3.7) die Summe der s üblichen Loglikelihoodfunktionen für die einfachen linearen Regressionsmodelle der Komponenten $j=1,\ldots,s$ mit den Daten (x_i,y_i), für die $\zeta(i)=j$ gilt. Also wird $\ln\hat{L}_n$ für gegebene $\hat{\zeta}(i)$, $i=1,\ldots,n$, durch die üblichen ML-Schätzer $(\hat{\beta}_j,\hat{\sigma}_j^2)$, d.h. die KQ-Schätzer für β_j und

$$\hat{\sigma}_j^2 := \frac{\sum_{\hat{\zeta}(i)=j}(y_i-\hat{\beta}_j'x_i)^2}{\hat{n}_j}, \quad \hat{n}_j := \sum_{i=1}^{n} 1(\hat{\zeta}(i)=j),\ j=1,\ldots,s, \tag{3.10}$$

maximiert. Damit vereinfacht sich die geschätzte Loglikelihoodfunktion, denn wegen (3.8) ist $\hat{r}(j) = \hat{n}_j$ für $j = 1, \ldots, s$.

Beginnt man also mit einer Startpartition $\hat{\zeta}(i)$, $i = 1, \ldots, n$ und iteriert dann abwechselnd die $(\hat{\beta}_j, \hat{\sigma}_j)$ gemäß (3.10) und die $\hat{\zeta}(i)$ gemäß (3.9), so wird in jedem Schritt $\ln \hat{L}_n$ vergrößert, bis die Iteration, da die $\hat{\zeta}(i)$ diskret sind, nach endlich vielen Schritten ein lokales Maximum erreicht hat. Das macht die Berechnung des ML-Schätzers im Fixed Partition Model nach meiner Erfahrung wesentlich schneller als im Mischmodell. Die Vorgehensweise ist eine einfache Variation des EM-Algorithmus aus Abschnitt 3.3.

Es kann allerdings der Fall $\dim(x_i : \hat{\zeta}(i) = j) < p + 1$ für ein j eintreten. Dann ist der KQ-Schätzer nicht mehr berechenbar. Außerdem muß bei der Berechnung der Fall $\hat{\sigma}_j^2 = 0$, also $\hat{L}_n = \infty$ ausgeschlossen werden. Dieser Fall tritt insbesondere ein, falls $\hat{n}_j \leq p + 1$.

Unter der Voraussetzung $\sigma_1^2 = \ldots = \sigma_s^2$ ist der ML-Schätzer für die $\zeta(i)$ und β_j der KQ-Schätzer aus (3.2), wie man (3.7) entnehmen kann.

Zur Schätzung der Anzahl der Cluster gibt es auch im Lokationsproblem kaum Anhaltspunkte. Banfield und Raftery (1993) schlagen einen Bayes'schen Ansatz vor, der aber die Vorgabe einer a priori-Verteilung über $\{\gamma : I \mapsto I\!R^{p+1} \times I\!R_0^+\}$ benötigt.

Nach meiner Erfahrung sind das AIC (3.4) und das BIC (3.1) für diese Situation nicht geeignet, da die Loglikelihoodfunktion mit n schneller steigt als $\ln n$, denn im Fixed Partition Model steigt die Anzahl der Parameter mit n. Für die Simulationen habe ich daher das BIC wie folgt modifiziert:

$$\ln L_n(s) - \frac{1}{2} k(s) \ln n - 0.7\, sn \stackrel{!}{=} \max_{s \in I\!N} \tag{3.11}$$

Dabei sei wieder $\ln L_n(s)$ die Loglikelihoodfunktion im durch s definierten Modell an der Stelle des ML-Schätzers, d.h. $\ln L_n(s) := \ln L_n(s, \hat{\gamma}_{ML}, \mathbf{Z})$. $k(s) = (p+2)s$ sei die Anzahl der zu schätzenden Regressions- und Skalenparameter. Zusätzlich werden n weitere Parameter $\zeta(i)$ mit Wertebereich $\{1, \ldots, s\}$ geschätzt. Die Simulationsergebnisse deuten darauf hin, daß der Summand $-0.7\, sn$ eine sinnvolle Wahl ist, um damit umzugehen. Genaueres dazu findet sich in Teil IV.

Bemerkung 3.4 *Marriott (1975) weist darauf hin, daß bei bekanntem s die entsprechenden ML-Schätzer für den Lokationsfall bei multivariaten Normalverteilungen nicht konsistent, sondern systematisch verzerrt sind. Die Argumentation von Marriott ist auch auf den Regressionsclusterfall anwendbar. Sie beruht darauf, daß die ML-Schätzungen für Lage- bzw. Regressionsparameter und Skala für jeden Cluster die einfache ML-Schätzung mit den dem Cluster zugehörigen Punkten ist, wenn die Zuordnung der Punkte zu den Clustern gegeben ist (siehe oben). Die Zuordnungsschätzung nach (3.9) teilt den $I\!R^{p+1}$ in s Bereiche auf. Damit ist die ML-Parameterschätzung genaugenommen die Schätzung der Parameter eines Modells, dessen Störterm abgeschnitten normalverteilt ist, eingeschränkt auf den Bereich der Punkte, die dem entsprechenden Cluster zugeordnet werden. Daraus ergibt sich eine Unterschätzung der Störskala und - wenn der Zuordnungsbereich nicht symmetrisch um die Regressionshyperebene ist - auch eine Verzerrung der Regressionsparameter.*

Bemerkung 3.5 *Sei $\mathbf{Z}^D$ definiert wie in Bemerkung 3.3. Dann gilt wegen (3.5):*

$$\ln L_n(s, \gamma(1), \ldots, \gamma(n), \mathbf{Z}) - n \ln a = \ln L_n(s, \gamma^*(1), \ldots, \gamma^*(n), \mathbf{Z}^D),$$
$$\text{wobei } \gamma^*(i) := ((\Gamma^{-1})'(a\beta(i) + b), a^2\sigma(i)^2) \; \forall i = 1, \ldots, n.$$

Damit ist der Fixed Partition ML-Schätzer linear äquivariant. Das heißt: Ist

$$\gamma \in \left(I\!R^{p+1} \times I\!R^{+}\right)^{n}$$

(lokale) Minimalstelle von $\ln L_n(s, \bullet, \mathbf{Z})$ *unter der Nebenbedingung* $|\gamma(\{1, \ldots, n\})| = s$, *so ist das oben definierte* γ^* *(lokale) Minimalstelle von* $\ln L_n(s, \bullet, \mathbf{Z}^D)$ *unter derselben Nebenbedingung. Weiterhin ist die Schätzung von s nach (3.11) invariant unter D gemäß (2.2) analog zu Bemerkung 3.3 und entsprechend auch nach dem AIC und BIC.*

3.5 Alternative Ansätze

3.5.1 Robuste Regression

Die Verfahren, die nun behandelt werden, wurden nicht speziell zur Analyse der Modelle in Abschnitt 2 entwickelt. Es handelt sich um heuristisch begründete Datenanalyse-Verfahren, deren statistische Eigenschaften nicht erforscht sind und die andere Ziele haben als die reine Parameterschätzung.

Ich habe meine Arbeit an den Problemstellungen dieser Dissertation begonnen, indem ich mich dafür interessiert habe, ob lokale Minima robuster Regressionsschätzer dazu brauchbar sind, Cluster linearer Regression zu finden. Zu diesem Thema gibt es meines Wissens bislang nur eine Arbeit von Morgenthaler (1990).

Es gibt zwischen robuster Statistik und Clusteranalyse folgenden Zusammenhang: Die robuste Statistik bemüht sich um Schätzer, die möglichst nicht von Ausreißern beeinflußt werden, während in der Clusteranalyse die Punkte eines Clusters im Verhältnis zu den anderen Clustern Ausreißer sind, sofern es sich um gut getrennte Cluster handelt. In einer Situation mit zwei Clustern ist also die Problemstellung, die Parameter des größeren Clusters zu finden, äquivalent zu einer robusten Schätzung für den Gesamtdatensatz, die nicht davon beeinflußt wird, daß bis zur Hälfte der Daten aus einem anderen Modell stammen, nämlich dem kleineren Cluster. Robuste Regressionsschätzer sind meistens als globale Minima einer Zielfunktion definiert. Die Parameter, die die Zielfunktion global minimieren, hängen von mindestens der Hälfte der Punkte ab, denn es ist das Ziel robuster Schätzung, mindestens die Hälfte der Punkte gut anzupassen. Robuste Schätzer sind also keine sinnvollen Parameterschätzer für kleinere Cluster. Man kann sich aber überlegen, ob nicht lokale Minima die Existenz von kleineren Clustern indizieren könnten. Formaler:

Sei $\mathbf{Z}$ der Datensatz. Dann ist ein M-Schätzer mit allgemeiner Skala für den Regressionsparameter β eines homogenen Regressionsmodells definiert gemäß

$$\hat{\beta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho(r_i(\theta)), \text{ wobei } r_i(\theta) := \frac{y_i - \theta' x_i}{\hat{\sigma}(\mathbf{Z})}, \ \rho : I\!R \mapsto I\!R_0^+. \tag{3.12}$$

Dabei ist $\hat{\sigma}(\mathbf{Z})$ ein Skalenschätzer. Weiterhin werden üblicherweise einige Voraussetzungen an ρ gemacht, von denen insbesondere Achsensymmetrie um 0, Beschränktheit und monotones Wachstum für positive Argumente wichtig sind, in unserem Zusammenhang sogar streng monotones Wachstum bis zu einem Argument $a > 0$ und dann Konstanz. Weiterhin kann man verschiedene Glattheitsforderungen stellen. In diese Klasse fallen:

- S-Schätzer (Rousseeuw und Yohai (1988)). Dabei gilt für $\hat{\sigma}(\mathbf{Z}, \hat{\beta})$ bei gegebenem Schätzer des Regressionsparameters $\hat{\beta}$:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i - \hat{\beta}' x_i}{\hat{\sigma}(\mathbf{Z}, \hat{\beta})}\right) \stackrel{!}{=} \frac{\max\rho}{2}. \tag{3.13}$$

  Der S-Regressionsschätzer $\hat{\beta}_S$ ist nun definiert als Minimalstelle von $\hat{\sigma}(\mathbf{Z}, \hat{\beta})$ und erfüllt daher (3.12) für gegebenes $\hat{\sigma}(\mathbf{Z}, \hat{\beta}_S)$ aufgrund der Monotonie von ρ.

- MM-Schätzer (Yohai (1988)). In diesem Fall ist $\hat{\sigma}(\mathbf{Z})$ ein M-Skalenschätzer. Das kann zum Beispiel der in (3.13) definierte S-Skalenschätzer mit einer anderen als der für den MM-Regressionsschätzer verwendeten ρ-Funktion sein.

- „Redescending" (wiederabsteigende) M-Schätzer (Morgenthaler (1990)). In diesem Fall ist $\hat{\sigma}(\mathbf{Z}) = k$ konstant vorgegeben. Die Schätzer heißen „wiederabsteigend", weil 0 die Ableitung von ρ außerhalb von $[-a, a]$ ist. $\rho'(r_i(\hat{\beta}_{i,t}))$ ist eng mit dem Einfluß der Punkte z_i auf die Regressionsschätzung verbunden.

Die Zielfunktion in (3.12) wird üblicherweise durch eine iterierte gewichtete KQ-Regression minimiert. Um ein globales Minimum zu finden, führt man die Iteration mehrfach von zufällig gewählten Startpunkten aus durch. Dabei erhält man als Nebenprodukt meistens mehrere lokale Minima, die eine Fixpunktgleichung der Form

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}(\hat{\beta})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\hat{\beta})y \tag{3.14}$$

erfüllen, wobei $\mathbf{W}(\hat{\beta})$ eine Diagonalmatrix mit den vom konkreten Schätzverfahren abhängigen Gewichten der Beobachtungen ist. Insbesondere hat eine Beobachtung z_i das Gewicht 0, wenn $|r_i| > a$, d.h. $\rho(r_i) = \max\rho$. Ich nenne solche Beobachtungen „vom Schätzverfahren als Ausreißer klassifizierte Daten".

Im Falle eines lokalen Minimums wurde also der Teil der Daten z_i mit $|r_i| \leq a$ im Sinne einer gewichteten KQ-Regression „gut" angepaßt. Die Idee liegt nahe, daß diese Daten zu einem Cluster gehören, also gemeinsam von einem Regressionsmodell erzeugt wurden, dessen Regressionsparameter durch die lokale Minimalstelle von (3.12) geschätzt wird. Man könnte nun alle Mengen von Punkten, zu denen es ein lokales Minimum gibt, welches sie nicht als Ausreißer klassifiziert, als „Cluster" definieren.

Diese Vorgehensweise unterscheidet sich prinzipiell von den bisher vorgestellten Schätzverfahren:

- Ein Cluster wäre hier „lokal" definiert. Damit meine ich: Die Lage seiner Punkte im Verhältnis zum restlichen Datensatz ist entscheidend, nicht aber ein globales Modell. Es gibt keine globale Modellannahme.

- Es wird keine Partition erzwungen. Das heißt: Es kann Teile des Datensatzes geben, die zu keinem Cluster gehören. Es kann Punkte geben, die zu mehreren Clustern gehören.

- Es wird keine Optimallösung erzwungen. Das heißt: Aus den gefundenen Clustern sind unter Umständen mehrere alternative Auswahlen von „relevanten" Clustern möglich, es kann alternative Interpretationen geben.

Ich halte die Existenz von Datenanalyse-Verfahren, die diese Eigenschaften erfüllen, für sehr wünschenswert. Wann immer die Modellvoraussetzungen nicht genau erfüllt sind, für die Schätzverfahren zur Verfügung stehen (zum Beispiel im Telefondatensatz), wann immer man mehr über seinem Datensatz wissen will als einen speziellen Parameter, gehen wertvolle Information durch die Anpassung eines global vorausgesetzten Modelles verloren.

Ich werde nun diskutieren, warum ich die Regressions-Clusteranalyse mit robusten Regressionsschätzern trotzdem außer in Spezialfällen nicht besonders sinnvoll finde. Die bisherigen Ideen werden aber in dieser Arbeit weiterhin präsent sein: Das Konzept des Fixpunktclusters, das ich in Teil II einführe, beruht darauf, die oben aufgezählten Eigenschaften zu erhalten und die folgenden Nachteile zu vermeiden:

- S-Schätzer sind für unsere Zwecke unbrauchbar, da $\hat{\sigma}(\mathbf{Z}, \hat{\beta}_S)$ und ρ dort so definiert sind, daß auch für jedes lokale Minimum weniger als die Hälfte der Punkte als Ausreißer klassifiziert werden: Sei $n^* > \frac{n}{2}$ die Anzahl der Punkte z_i mit $\rho(r_i(\hat{\beta}_S)) = \max \rho$. Dann ist im Widerspruch zu (3.13)

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{\beta}_S' x_i}{\hat{\sigma}(\mathbf{Z}, \hat{\beta}_S)}\right) \geq \frac{n^* \max \rho}{n}.$$

  Also kann man mit S-Schätzern keine Cluster finden, die weniger als die Hälfte der Beobachtungen enthalten.

- Das Problem bei MM-Schätzern ist ähnlich. Zwar erzwingt die Definition keine Cluster, die mehr als die Hälfte der Punkte enthalten, aber die Skalenschätzung $\hat{\sigma}(\mathbf{Z})$ beruht immer mindestens auf der Hälfte der Daten. Sie wird unsinnig, wenn es keinen gemeinsamen Cluster gibt, der diese Punkte enthält. Da obendrein dasselbe $\hat{\sigma}(\mathbf{Z})$ für jede Iteration verwendet wird, wird implizit vorausgesetzt, daß alle Cluster dieselbe Störvarianz haben. Das Verfahren führt zu unbrauchbaren Ergebnissen, wenn sich die Skalen der übrigen Cluster deutlich von derjenigen des größten Clusters unterscheiden.

- Im Falle der wiederabsteigenden M-Schätzer schlägt Morgenthaler (1990) vor, lokale Minima für eine endliche absteigende Folge $k_1 \geq \ldots > k_d > 0$ von Skalenparametern zu berechnen. Damit werden die Probleme der vorherigen Verfahren umgangen. Andererseits scheint mir der numerische Aufwand bei größeren p und n extrem groß zu sein (Morgenthaler führt zwei Datenbeispiele mit $p = 2, n = 11$ und 24 an). Die Menge der gefundenen lokalen Minima insbesondere bei kleinem k kann unüberschaubar werden.

Zum Schluß sei noch auf das einfache Verfahren von Curcic und Pierantoni (1995) hingewiesen. Sie empfehlen, den robusten LMS-Schätzer (siehe zum Beispiel Rousseeuw und Leroy (1988)) nach seiner Berechnung auch noch auf die Daten anzuwenden, die im ersten Durchgang als Ausreißer klassifiziert wurden. Um damit ein sinnvolles Ergebnis zu bekommen, ist natürlich aber wieder ein Cluster vonnöten, der mehr als die Hälfte der Punkte umfaßt.

3.5.2 Schwache Hierarchien

Hierarchische Clusteranalyseverfahren beruhen normalerweise auf einer Abstands- oder Ähnlichkeitsrelation zwischen den Daten. Der (euklidische) Abstand zwischen den Datenpunkten sagt im Regressionsfall aber nichts darüber aus, ob die Punkte von einer gemeinsamen Regressionsverteilung erzeugt wurden oder nicht. Durch $p+1$ Punkte, die beliebig weit voneinander entfernt sind, kann ja noch eine gemeinsame p- dimensionale Hyperebene gelegt werden. Wollte man lineare Regressionscluster hierarchisch analysieren, könnte man aber auf die Methode von Bandelt und Dress (1994) zurückgreifen. Gegeben sei eine d-variate Abstandsfunktion $\delta : X^d \mapsto I\!R_0^+$, $d \in I\!N$. Dann ist eine Menge $C \subseteq X$ ein Cluster, wenn

$$\forall c_1, \ldots, c_d \in C, b \in X - C : \delta(c_1, \ldots, c_d) < \max_{1 \le i \le d} \delta(c_1, \ldots, c_{i-1}, b, c_{i+1}, \ldots, c_d). \quad (3.15)$$

Bandelt und Dress (1994) zeigen, daß das dadurch induzierte Mengensystem eine „schwache Hierarchie" ist. Im linearen Regressionsfall könnte man nun für $d \ge p+2$

$$\delta(z_1, \ldots, z_d) := \frac{1}{d} \sum_{i=1}^{d} (y_i - \hat{\beta}'_{KQ} x_i)^2$$

definieren, wobei $\hat{\beta}_{KQ}$ der Kleinste-Quadrate-Regressionsschätzer für $z_1, \ldots, z_d$ sei[8]. δ wird also klein, sobald $z_1, \ldots, z_d$ gut durch eine gemeinsame Regressionsebene angepaßt werden können. Auch dieser Ansatz erzwingt keine Partition und ergibt einen „lokal" definierten Clusterbegriff. Ich habe ihn aber nicht weiterverfolgt, weil der Rechenaufwand, der benötigt wird, um die Mengen C zu finden, die (3.15) erfüllen, für $n \ge 50$ exzessiv ist. Systematische Simulationen sind daher derzeit noch unmöglich.

Für Daten, die man sich als eine Stichprobe aus einem statistischen Modell vorstellt, ist der Ansatz darüberhinaus nicht angemessen. Ähnlich wie beim „Single-Linkage-Clustering" müssen nämlich auch hier die Cluster scharf voneinander getrennt sein, um (3.15) zu erfüllen. Vereinzelte Punkte zwischen den dichtesten Bereichen der „anschaulichen" Cluster bewirken meistens, daß (3.15) von diesen Punktmengen nicht mehr erfüllt wird. Solche Punkte kommen aber automatisch vor, wenn man eine immer größere Stichprobe aus einer Verteilung zieht, deren Träger der gesamte $I\!R^{p+1}$ ist.

[8] Wie in Abschnitt 3.2 werden hier implizit gleiche Störvarianzen für alle Cluster vorausgesetzt.

4 Einführung: Identifizierbarkeit

Für Mischmodelle bedeutet „Identifizierbarkeit", daß dieselbe Mischverteilung nicht durch verschiedene vermischende Verteilungen (die J aus Modell 1 bzw. 3) auf einer gegebenen Menge von Verteilungen konstruiert werden kann. Eingeführt wurde das Identifizierbarkeitskonzept in diesem Zusammenhang von Teicher (1961).

Die Frage der Identifizierbarkeit ist von großer Bedeutung, wenn man innerhalb eines Modelles (zum Beispiel den Modellen aus Abschnitt 2) Parameter schätzen will. Ist nämlich ein Modell nicht identifizierbar, d.h. gibt es mehrere vermischende Verteilungen, die dieselbe Mischverteilung erzeugen, so ist konsistente Parameterschätzung nicht möglich: Auch asymptotisch kann nicht zwischen den durch die vermischenden Verteilungen definierten verschiedenen Parametern unterschieden werden.

In dieser Arbeit sind nur endliche Mischungen auf $\mathbb{R}^d$ von Interesse, d.h. die vermischende Verteilung hat endlichen Träger:

Definition 4.1 (Identifizierbarkeit: Endliche Mischmodelle) *Sei T ein Parameterraum mit einer σ–Algebra $\mathcal{T}$, die alle Einpunktverteilungen auf T enthält. $\mathcal{J}(T)$ sei wieder die Menge der Verteilungen auf $(T, \mathcal{T})$ mit endlichem Träger, $\mathcal{F} := \{F(\bullet, \theta) : \theta \in T\}$ sei eine Menge von Verteilungsfunktionen auf $\mathbb{R}^d$,*

$$C_{\mathcal{J}(T)} := \left\{ H_J : H_J(x) = \int_T F(x, \theta) dJ(\theta), \quad x \in \mathbb{R}^d, J \in \mathcal{J}(T) \right\}.$$

Ist nun $Q : \mathcal{J}(T) \mapsto C_{\mathcal{J}(T)}$ durch $Q(J) = H_J$ definiert, dann ist $C_{\mathcal{J}(T)}$ identifizierbar, falls Q bijektiv ist.

Bemerkung 4.2 *Auch hier noch einmal die alternative Schreibweise von Titterington, Smith und Makov (1985) (siehe Bemerkung 2.1): Mit der Notation aus Definition 4.1 sei*

$$C = \{H : H(x) = \sum_{i=1}^{s} \epsilon_i F(x, \theta_i), \quad \epsilon_i > 0, \quad \sum_{i=1}^{s} \epsilon_i = 1,$$
$$\theta_i \in T, \quad \forall i = 1, \ldots, s, \quad s \in \mathbb{N}, x \in \mathbb{R}^p\}.$$

Dann ist C identifizierbar, wenn für zwei beliebige Elemente $H, \hat{H}$ aus C, definiert durch

$$H = \sum_{i=1}^{s} \epsilon_i F(\bullet, \theta_i), \qquad \hat{H} = \sum_{i=1}^{\hat{s}} \hat{\epsilon}_i F(\bullet, \hat{\theta}_i),$$

θ_i, $i = 1, \ldots, s$ (bzw. $\hat{\theta}_i$, $i = 1, \ldots, \hat{s}$) paarweise verschieden, dann und nur dann $H = \hat{H}$ gilt, wenn $s = \hat{s}$ und es eine s–Permutation Π gibt, so daß

$$((\epsilon_1, \theta_1), \ldots, (\epsilon_s, \theta_s)) = ((\hat{\epsilon}_{\Pi(1)}, \hat{\theta}_{\Pi(1)}), \ldots, (\hat{\epsilon}_{\Pi(s)}, \hat{\theta}_{\Pi(s)})).$$

s entspricht dabei $|S(J)|$, ϵ_i entspricht $J\{\theta_i\}$ aus Definition 4.1.

Einige häufig verwendete Verteilungsklassen erzeugen identifizierbare Mischungen, zum Beispiel multivariate Normalverteilungen (Yakowitz and Spragins (1968)). Es gibt jedoch bislang keine Ergebnisse über lineare Regression. In Abschnitt 5 wird gezeigt, daß sich die Resultate für Normalverteilungen auch dann nicht einfach auf den Regressionsfall übertragen, wenn man einfache Überparametrisierung (d.h. Mischungskomponenten, deren Regressoren kollinear sind) ausschließt.

Der Identifizierbarkeitsbegriff aus Definition 4.1 reicht für die hier betrachteten Modelle nicht aus:

- Verteilungen wie in den Modellen 1 und 2 haben nicht die Form aus Definition 4.1. Produktverteilungen werden in dieser Definition nicht vorgesehen.
- Es kann (und wird) so sein, daß nur ein Teil der Parameter identifizierbar und daher auch konsistent schätzbar ist. Zum Beispiel sind in Modell 2 die Regressions- und Skalenparameter meistens identifizierbar (Satz 6.4), im Gegensatz zur Clusterzugehörigkeit der Punkte (Beispiel 5.3).

Für diese Zwecke wird der Identifizierbarkeitsbegriff nun verallgemeinert.

Definition 4.3 (Identifizierbarkeit) *Sei Ω ein beliebiger Parameterraum, $\mathcal{P}$ eine Menge von Verteilungen,*

$$C_\Omega = (F_\omega)_{\omega\in\Omega} \in \mathcal{P}^\Omega,$$

„$\sim$" eine Äquivalenzrelation auf Ω. Dann heißt C_Ω identifizierbar bzgl. „$\sim$", falls

$$\forall \omega, \hat{\omega} \in \Omega: \qquad F_\omega = F_{\hat{\omega}} \Leftrightarrow \omega \sim \hat{\omega}.$$

Definition 4.4 (Teilweise Identifizierbarkeit) *Mit denselben Bezeichnungen heißt C_Ω teilweise identifizierbar bzgl. „$\sim$", falls*

$$\forall \omega, \hat{\omega} \in \Omega: \qquad F_\omega = F_{\hat{\omega}} \Rightarrow \omega \sim \hat{\omega}.$$

Bemerkung 4.5 *C_Ω ist hier ein geordnetes Tupel und keine Menge, da die konkrete Zuordnung $\omega \mapsto F_\omega$ wichtig für die Identifizierbarkeit ist. Wäre $C_\Omega = \{F_\omega : \omega \in \Omega\}$, so könnten zwei unterschiedliche Zuordnungen definiert werden, so daß C_Ω bzgl. derselben Äquivalenzrelation identifizierbar und nicht identifizierbar wäre.*

Bemerkung 4.6 *Mit den Bezeichnungen aus Definition 4.1 definiere*

$$J \sim_T \hat{J} :\Leftrightarrow J = \hat{J} \quad \forall J, \hat{J} \in \mathcal{J}(T).$$

Dann ist „$\sim_T$" eine Äquivalenzrelation und $C_{\mathcal{J}(T)}$ ist genau dann identifizierbar bzgl. „$\sim_T$", wenn $C_{\mathcal{J}(T)}$ im Sinne von Definition 4.1 identifizierbar ist.

Bemerkung 4.7 *Sei C_Ω identifizierbar bzgl. „$\sim$". Ist $\Omega_1 \subset \Omega$ und „$\sim_1$" die Einschränkung von „$\sim$" auf Ω_1, so ist C_{Ω_1} trivialerweise auch identifizierbar bzgl. „$\sim_1$". Dieselbe Inklusion gilt für teilweise Identifizierbarkeit. Das bedeutet zum Beispiel, daß aus der (teilweisen) Identifizierbarkeit von Mischmodellen (nach Bemerkung 4.6) immer auch die (teilweise) Identifizierbarkeit der Modelle mit fest vorgegebener Zahl von Mischkomponenten $|S(J)| = s$ folgt.*

Bemerkung 4.8 *Prakasa Rao (1992) gibt einen Überblick über Identifizierbarkeitsprobleme in allgemeineren Situationen als Mischungen und Fixed-Partition-Clustern. Die Definitionen 4.3 und 4.4 würden alle diese Probleme abdecken, sofern man sie noch auf Identifizierbarkeit einzelner Äquivalenzklassen verallgemeinern würde. Insbesondere ist meine Definition von teilweiser Identifizierbarkeit eine Verallgemeinerung der „partial identifiability" bei Prakasa Rao auf S. 149.*

Die Definitionen werden nun auf die Modelle aus Abschnitt 2 angewendet.

Beispiel 4.9 (Modell 1) *Für festes $\tilde{x} = (x_i)_{i \in I}$ sei*

$$C_{\mathcal{J}(T_f)} = \left(F_{\tilde{x},J} : F_{\tilde{x},J} = \bigotimes_{i \in I} F_{x_i,J} \right)_{J \in \mathcal{J}(T_f)},$$

wobei $F_{x_i,J}$ wie in Modell 1 definiert sei. Weiterhin sei

$$J \sim_f \hat{J} :\Leftrightarrow J = \hat{J} \quad \forall J, \hat{J} \in \mathcal{J}(T_f).$$

Beispiel 4.10 (Modell 2) *Für festes $\tilde{x} = (x_i)_{i \in I}$ sei*

$$\Omega_p := \left\{ \gamma : I \mapsto \mathbb{R}^{p+1} \times \mathbb{R}_0^+, \quad |\gamma(I)| < \infty \right\},$$

$$C_{\Omega_p} = \left(F_{\tilde{x},\gamma} : F_{\tilde{x},\gamma} = \bigotimes_{i \in I} F_{x_i,\gamma(i)} \right)_{\gamma \in \Omega_p},$$

wobei die $F_{x_i,\gamma(i)}$, $i \in I$ wie in Modell 2 definiert seien.
Für $\gamma, \hat{\gamma} \in \Omega_p$ definiere nun

$$\gamma \sim_p \hat{\gamma} :\Leftrightarrow \gamma = \hat{\gamma}.$$

Beispiel 5.3 wird zeigen, daß C_{Ω_p} bzgl. „$\sim_p$" nicht identifizierbar ist. Es ist aber unter Umständen möglich, die Regressions- und Skalenparameter ohne die Clusterzugehörigkeit teilweise zu identifizieren. Dafür sei

$$\gamma \sim_{p1} \hat{\gamma} :\Leftrightarrow \gamma(I) = \hat{\gamma}(I).$$

Beispiel 4.11 (Modell 3) *Sei $C_{\mathcal{J}(T_s)} = (F_J)_{J \in \mathcal{J}(T_s)}$, F_J und T_s (versehen mit der σ–Algebra aus Bemerkung 2.3) wie in Modell 3, „$\sim_{T_s}$" definiert wie „$\sim$" in Bemerkung 4.6.*

Mit dieser Definition kann es unterschiedliche Mischungskomponenten geben, die dieselben Regressions- und Skalenparameter haben und sich nur in der Regressorenverteilung unterscheiden. Das verursacht Identifizierbarkeitsprobleme (Beispiel 5.1) und ist unter Umständen inhaltlich nicht angemessen: Wenn man unter einem „Regressionscluster" eine Menge von Datenpunkten versteht, die von Verteilungen mit gemeinsamen Regressions- und Skalenparametern generiert wurden, und man die Analyse des Mischmodells zu Zwecken der Clusteranalyse benutzt, dann sollte ein Cluster nicht mehreren Mischungskomponenten entsprechen. Statt $\mathcal{J}(T_s)$ kann man dann

$$\Omega_s := \{ J \in \mathcal{J}(T_s) : (\beta, \sigma^2, G) \in S(J), G \neq \hat{G} \Rightarrow (\beta, \sigma^2, \hat{G}) \notin S(J) \}$$

verwenden. Auch hier gelte wieder

$$J \sim_s \hat{J} :\Leftrightarrow J = \hat{J} \quad \forall J, \hat{J} \in \Omega_s.$$

Unter Umständen (Bemerkung 4.12) wird $C_{\mathcal{J}(T_s)}$ durch die Ersetzung von $\mathcal{J}(T_s)$ durch Ω_s nicht eingeschränkt.

Satz 6.8 bringt die Identifizierbarkeit von C_{Ω_s} bzgl. „$\sim_s$" unter einer Voraussetzung an $\mathcal{G}$. Falls man nicht an der Identifikation der Regressorenverteilung interessiert ist und die Voraussetzung aus Satz 6.8 nicht erfüllt ist, kann man C_{Ω_s} noch auf teilweise Identifizierbarkeit untersuchen. Dafür sei

$$J \sim_{s0} \hat{J} :\Leftrightarrow \forall (\beta, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}: \; J\{(\beta, \sigma^2, G) : G \in \mathcal{G}\} = \hat{J}\{(\beta, \sigma^2, G) : G \in \mathcal{G}\}.$$

In Beispiel 5.2 wird sich jedoch zeigen, daß auf Ω_s eine noch schwächere Äquivalenzrelation „$\sim_{s1}$" gebraucht wird, um teilweise Identifizierbarkeit von C_{Ω_s} zu zeigen:

$$J \sim_{s1} \hat{J} :\Leftrightarrow$$
$$\Leftrightarrow \{(\beta, \sigma^2) : (\beta, \sigma^2, G) \in S(J)\} = \{(\beta, \sigma^2) : (\beta, \sigma^2, G) \in S(\hat{J})\}.$$

Bemerkung 4.12 *Enthält $\mathcal{G}$ alle endlichen Mischungen seiner Elemente, so gilt*

$$C_0 := \{F_J : J \in \mathcal{J}(T_s)\} = \{F_J : J \in \Omega_s\} =: C_*.$$

Beweis: „$\supseteq$" ist klar. „$\subseteq$" ist zu zeigen. Sei $F_J \in C_0$, $J_0 := J$, $J_0 \notin \Omega_s$, so daß $\exists \beta, \sigma^2, G_1, G_2$ mit

$$G_1 \neq G_2, \quad J_0\{(\beta, \sigma^2, G_1)\} = \epsilon_1, \; J_0\{(\beta, \sigma^2, G_2)\} = \epsilon_2, \; \epsilon_1, \epsilon_2 > 0. \tag{4.1}$$

Setze $G_3 := \frac{\epsilon_1 G_1 + \epsilon_2 G_2}{\epsilon_1 + \epsilon_2}$. Definiere nun J_* gemäß

$$J_*\{\theta\} = J_0\{\theta\} \; \forall \theta \in S(J_0) \setminus \{(\beta, \sigma^2, G_1), (\beta, \sigma^2, G_2), (\beta, \sigma^2, G_3)\} =: S_-(J_0),$$
$$J_*\{(\beta, \sigma^2, G_1)\} := J_*\{(\beta, \sigma^2, G_2)\} := 0,$$
$$J_*\{(\beta, \sigma^2, G_3)\} := J_0\{(\beta, \sigma^2, G_3)\} + \epsilon_1 + \epsilon_2.$$

Es folgt nun mit den Bezeichnungen aus Modell 3 aus den Definitionen von F_J und $F(\bullet, \theta)$:

$$\begin{aligned} F_J(x,y) = F_{J_0}(x,y) &= \epsilon_1 F(x,y,(\beta,\sigma^2,G_1)) + \epsilon_2 F(x,y,(\beta,\sigma^2,G_2)) + \\ &+ J_0\{(\beta,\sigma^2,G_3)\}F(x,y,(\beta,\sigma^2,G_3)) + \sum_{\theta \in S_-(J_0)} J_0\{\theta\}F(x,y,\theta) = \\ &= (\epsilon_1+\epsilon_2)\left(\tfrac{\epsilon_1}{\epsilon_1+\epsilon_2}F(x,y,(\beta,\sigma^2,G_1)) + \tfrac{\epsilon_2}{\epsilon_1+\epsilon_2}F(x,y,(\beta,\sigma^2,G_2))\right) + \\ &+ J_0\{(\beta,\sigma^2,G_3)\}F(x,y,(\beta,\sigma^2,G_3)) + \sum_{\theta \in S_-(J_0)} J_0\{\theta\}F(x,y,\theta) = \\ &= (\epsilon_1+\epsilon_2+J_0\{(\beta,\sigma^2,G_3)\})F(x,y,(\beta,\sigma^2,G_3)) + \sum_{\theta \in S_-(J_0)} J_0\{\theta\}F(x,y,\theta) = \\ &= F_{J_*}(x,y). \end{aligned}$$

Nach Voraussetzung ist $G_3 \in \mathcal{G}$. Daher ist nun ist entweder $J_* \in \Omega_s$ und also $F_J \in C_*$ oder es gibt weitere $\beta, \sigma^2, G_1, G_2$, so daß (4.1) auch für $J_0 := J_*$ gilt. In diesem Fall wende man dasselbe Verfahren noch einmal an. Nach einer endlichen Anzahl von solchen Schritten ($|S(J)|$ ist endlich und $|S(J_*)| \leq |S(J)|$) ist $J_* \in \Omega_s$. Also gilt $F_J \in C_*$, was zu zeigen war.

5 Beispiele für Nicht-Identifizierbarkeit

In diesem Abschnitt wird gezeigt, daß Identifizierbarkeit von Regressionsmischungen mit normalverteiltem Störterm nicht automatisch aus der Identifizierbarkeit für Normalverteilungen folgt, wie zum Beispiel DeSarbo und Cron (1988) behaupten. Man kann aus den Beispielen ableiten, welche Parameter unter welchen Umständen nicht konsistent schätzbar sind. Zwar folgt aus den Beispielen nicht, daß die Voraussetzungen, die in Abschnitt 6 für Identifizierbarkeit gegeben werden, so allgemein wie möglich sind, aber es wird doch gezeigt, daß zumindest vergleichbare Voraussetzungen nötig sind, um Identifizierbarkeit zu erhalten. Es gibt zwei wesentliche Gründe für die Nicht-Identifizierbarkeit linearer Regressionsmischungen:

- Nicht-Identifizierbarkeit wegen Problemen mit den Regressoren: Nicht-Identifizierbarkeit der Regressorenverteilung in Modell 3 oder Nicht-Identifizierbarkeit der Clusterzuordnung in Modell 2, weil der Regressor so liegt, daß die Zuordnung nicht eindeutig ist.
- Nicht-Identifizierbarkeit wegen zu vieler Cluster für zu wenig Regressoren. Es wird sich zeigen, daß das auch passieren kann, wenn für jeden einzelnen Cluster genügend Regressoren vorhanden sind, d.h. kein Cluster für sich überparametrisiert ist.

Beispiel 5.1 (Regressoren nicht identifizierbar) *Es gelten die Bezeichnungen aus Modell 3 und Beispiel 4.11. Es enthalte nun $\mathcal{G}$ nicht-identifizierbare endliche Mischungen seiner Elemente, also*

$$G = \int_{\mathcal{G}} P dJ(P) = \int_{\mathcal{G}} P d\hat{J}(P) = \hat{G}, \quad J \neq \hat{J} \in \mathcal{J}(\mathcal{G}).$$

Dann ist $C_{\mathcal{J}(T_s)}$ nicht identifizierbar bzgl. „$\sim_{T_s}$“: Gegeben (β_0, σ_0^2) definiere

$$K\{(\beta_0, \sigma_0^2, G)\} := J\{G\} \quad \forall G \in \mathcal{G}, \tag{5.1}$$

$$K\{(\beta, \sigma^2, G)\} := 0 \quad \forall (\beta, \sigma^2) \neq (\beta_0, \sigma_0^2), \tag{5.2}$$

$\hat{K}$ mit $\hat{G}$ statt G. Dann ist $K \neq \hat{K}$, aber $F_K = F_{\hat{K}}$.

Dieser Fall tritt zum Beispiel ein, wenn $\mathcal{G}$ die empirischen Verteilungen enthält, da jede empirische Verteilung (bis auf die Einpunktverteilungen) eine Mischung anderer empirischer Verteilungen ist. Um dieses Problem zu vermeiden, kann $\mathcal{J}(T_s)$ durch Ω_s ersetzt werden.

Beispiel 5.2 (Positive Masse auf kollinearen Regressoren) *Wenn $\mathcal{G}$ Verteilungen enthält, unter denen eine $p-1$-dimensionale Hyperebene*

$$H_\alpha := \{x^- \in \mathbb{R}^p : x'\alpha = 0, \quad \alpha \in \mathbb{R}^{p+1} \setminus \{0\}\}$$

eine positive Wahrscheinlichkeit hat, ist C_{Ω_s} im allgemeinen nicht identifizierbar bzgl. „$\sim_s$“:

Für $x^- \in H_\alpha$ gilt $x'\beta = x'(\beta+\alpha)$. Im Fall $p=1$ zeigt Abbildung 4 diese Situation. Sei nun $G_1 \in \mathcal{G}$, wobei

$$\xi := G_1(H_\alpha) > 0, \quad G_H(B) := G_1(B|H_\alpha) \forall B \in I\!B^p,$$
$$G_2 := \tfrac{1}{1-\xi}(G_1 - \xi G_H) \text{ (muß für dieses Gegenbeispiel } \in \mathcal{G} \text{ sein)},$$
$$J := \tfrac{1}{2-\xi}\delta_{(\beta,\sigma^2,G_1)} + \tfrac{1-\xi}{2-\xi}\delta_{(\beta+\alpha,\sigma^2,G_2)},$$
$$\hat{J} := \tfrac{1-\xi}{2-\xi}\delta_{(\beta,\sigma^2,G_2)} + \tfrac{1}{2-\xi}\delta_{(\beta+\alpha,\sigma^2,G_1)} \Rightarrow$$
$$F_J(x,y) = \tfrac{1}{2-\xi}\int 1(t \le x)\Phi_{0,\sigma^2}(y - t'\beta)dG_1(t) +$$
$$+\tfrac{1-\xi}{2-\xi}\int 1(t \le x)\Phi_{0,\sigma^2}(y - t'(\beta+\alpha))dG_2(t) =$$
$$= \tfrac{1}{2-\xi}\int 1(t \le x)\Phi_{0,\sigma^2}(y - t'\beta)d[(1-\xi)G_2 + \xi G_H](t) +$$
$$+\tfrac{1-\xi}{2-\xi}\int 1(t \le x)\Phi_{0,\sigma^2}(y - t'(\beta+\alpha))d\left[\tfrac{1}{1-\xi}(G_1 - \xi G_H)\right](t) =$$
$$= \tfrac{1-\xi}{2-\xi}\int 1(t \le x)\Phi_{0,\sigma^2}(y - t'\beta)dG_2(t) +$$
$$+\tfrac{1}{2-\xi}\int 1(t \le x)\Phi_{0,\sigma^2}(y - t'(\beta+\alpha))dG_1(t) = F_{\hat{J}}(x,y), \text{ da}$$
$$(y - t'\beta) = (y - t'(\beta+\alpha)) \ [G_H].$$

Weiterhin ist $J \not\sim_s \hat{J}$, $J \not\sim_{s0} \hat{J}$. Daher ist C_{Ω_s} nicht identifizierbar bzgl. „$\sim_s$" und nicht einmal teilweise identifizierbar bzgl. „$\sim_{s0}$"; d.h. auch die Anteile der Mischungskomponenten sind nicht identifizierbar.

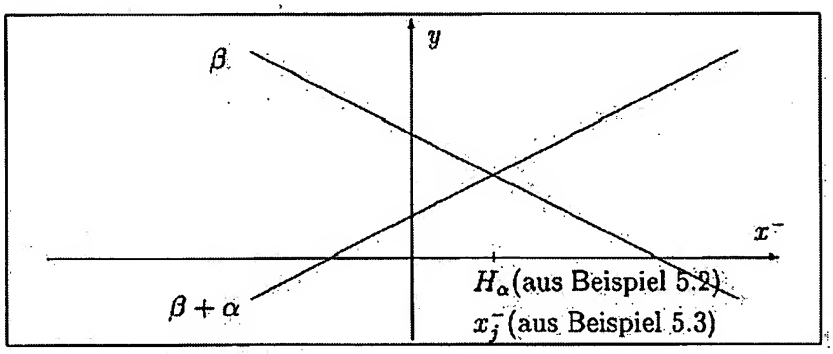


Abbildung 4: Clusterzuordnung nicht eindeutig

In Modell 2 sind aus ähnlichen Gründen die Clusterzuordnungen der Punkte nicht identifizierbar:

Beispiel 5.3 (Clusterzuordnung in Modell 2) *C_{Ω_p} ist nicht identifizierbar bzgl. „$\sim_p$":*
Sei $j \in I$. Wähle nun $\alpha \in I\!R^{p+1} \setminus \{0\}$ so, daß $x_j'\alpha = 0$. Sei

$$\gamma(I) = \{(\beta,\sigma^2), (\beta+\alpha,\sigma^2)\}.$$

Dann ist $F_{z_j,\gamma}$ und damit die gemeinsame Verteilung der $(y_i)_{i\in I}$ gleich für $\gamma(j) = (\beta,\sigma^2)$ und $\gamma(j) = (\beta+\alpha,\sigma^2)$.

In den folgenden Beispielen sind auch die Regressionsparameter nicht mehr identifizierbar. Natürlich ist

$$\forall(\beta,\sigma^2) \in \gamma(I): \dim\langle\{x_i : \gamma(i) = (\beta,\sigma^2)\}\rangle = p+1$$

notwendige Bedingung für Identifizierbarkeit in Modell 2. Analog muß auch in Modell 1 und Modell 3 die Identifizierbarkeit des Regressionsparameters jedes einzelnen Clusters gesichert werden. Es muß also ausgeschlossen werden, daß die Regressoren x^- für einen einzelnen Cluster auf einer gemeinsamen $p-1$-dimensionalen Hyperebene des $\mathbb{R}^p$ liegen. Das reicht jedoch nicht aus. Auch falls die Anzahl der $p-1$-dimensionalen Hyperebenen, die man benötigt, um die Regressorenpunkte für jeden einzelnen Cluster abzudecken, beliebig hoch ist, lassen sich noch Beispiele für Nicht-Identifizierbarkeit konstruieren.

Beispiel 5.4 (Zu viele Cluster: Gitterstruktur) *Sei $p = 1$, $I = \{1,\dots,n\}$, $n = s^2$, $s := |\gamma(I)|$, $x_i = (i,1)$. Entsprechende Situationen sind auch für andere, nicht äquidistante Regressorenkonstellationen konstruierbar. Es gelten die Bezeichnungen aus Modell 2 bzw. Beispiel 4.10. Nun sei*

$$\gamma(i) = (0, js, \sigma^2) :\Leftrightarrow i \in \{(j-1)s+1,\dots,js\},\ j = 1,\dots,s,$$

d.h. Steigungsparameter 0, Achsenabschnitt js, und

$$\hat{\gamma}(i) = (1, s-j, \sigma^2) :\Leftrightarrow i \in \{(k-1)s+j : k \in \{1,\dots,s\}\},\ j = 1,\dots,s.$$

Offenbar ist $\gamma \not\sim_{p1} \hat{\gamma}$. Alle $i \in I$ sind wie folgt eindeutig darstellbar: $i = (k-1)s+l$, $k \in \{1,\dots,s\}$, $l \in \{1,\dots,s\}$. Damit ist

$$F_{x_i,\gamma(i)} = \Phi_{0,\sigma^2}(y_i - ks) = \Phi_{0,\sigma^2}(y_i - [((k-1)s+l) * 1 + s - l]) = F_{x_i,\hat{\gamma}(i)}.$$

Also ist die gemeinsame Verteilung der $(y_i)_{i\in I}$ für γ und $\hat{\gamma}$ gleich und also C_{Ω_p} nicht teilweise identifizierbar bzgl. „$\sim_{p1}$".

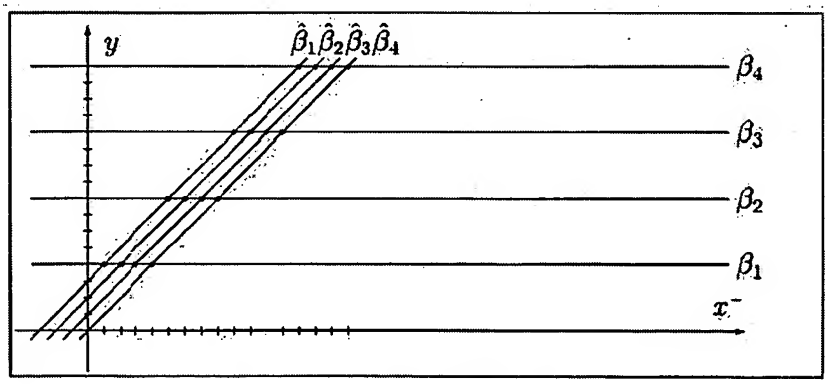


Abbildung 5: Gitterstruktur

Für $p > 1$ können anstatt eines einzelnen x_i^- beliebig viele Punkte auf einer gemeinsamen $p-1$-dimensionalen Hyperebene gewählt werden, so daß man Beispiele mit s Clustern und beliebig mehr als s^2 Punkten bekommt.

In Modell 1 ist jede Mischungskomponente in der Verteilung von y_i für jeden Regressor x_i vertreten. Das macht die Konstruktion von Gegenbeispielen etwas schwieriger. Die Anzahl der Regressoren n ist hier gleichzeitig die Anzahl der Punkte pro Mischungskomponente. Das heißt, daß eine vergleichbare Situation mit $|S(J)|$ statt wie oben s^2 Regressoren konstruiert werden muß.

Beispiel 5.5 (Regressionsparameter nicht identifizierbar in Modell 1) *Sei $p = 1$, $n = 3$, $|S(J)| = 3$. Beispiele mit größeren n, $|S(J)|$ sind möglich, aber viel komplizierter zu konstruieren. Für $p > 1$ lassen sich wie im vorigen Beispiel die Regressoren durch beliebig viele Punkte auf parallelen $p-1$-dimensionalen Hyperebenen ersetzen.*

Sei $x_1 = (0,1), x_2 = (1,1), x_3 = (2,1)$, $\sigma^2 \geq 0$ fest. Sei J definiert gemäß

$$j = 1,2,3: \ \sigma_j^2 = \sigma^2, \ \beta_1 = (\tfrac{3}{2}, 0), \ \beta_2 = (0,1), \ \beta_3 = (0,2),$$
$$S(J) := \{(\beta_j, \sigma_j^2) : j = 1,2,3\}, \ J\{(\beta_j, \sigma_j^2)\} := \tfrac{1}{3}$$

und $\hat{J} \not\sim_f J$ gemäß

$$j = 1,2,3: \ \hat{\sigma}_j^2 = \sigma^2, \ \hat{\beta}_1 = (-\tfrac{1}{2}, 2), \ \hat{\beta}_2 = (1,1), \ \hat{\beta}_3 = (1,0),$$
$$S(\hat{J}) := \{(\hat{\beta}_j, \hat{\sigma}_j^2) : j = 1,2,3\}, \ \hat{J}\{(\beta_j, \sigma_j^2)\} := \tfrac{1}{3}.$$

Dann ist die gemeinsame Verteilung der $(y_i)_{i\in\{1,2,3\}}$ in beiden Fällen das unabhängige Produkt der Verteilungen

$$F_{x_1} = \tfrac{1}{3}\mathcal{N}_{(0,\sigma^2)} + \tfrac{1}{3}\mathcal{N}_{(1,\sigma^2)} + \tfrac{1}{3}\mathcal{N}_{(2,\sigma^2)},$$
$$F_{x_2} = \tfrac{1}{3}\mathcal{N}_{(1,\sigma^2)} + \tfrac{1}{3}\mathcal{N}_{(\frac{3}{2},\sigma^2)} + \tfrac{1}{3}\mathcal{N}_{(2,\sigma^2)},$$
$$F_{x_3} = \tfrac{1}{3}\mathcal{N}_{(1,\sigma^2)} + \tfrac{1}{3}\mathcal{N}_{(2,\sigma^2)} + \tfrac{1}{3}\mathcal{N}_{(3,\sigma^2)}.$$

Also ist $C_{J(T_J)}$ in diesem Fall nicht identifizierbar bzgl. „$\sim_f$".

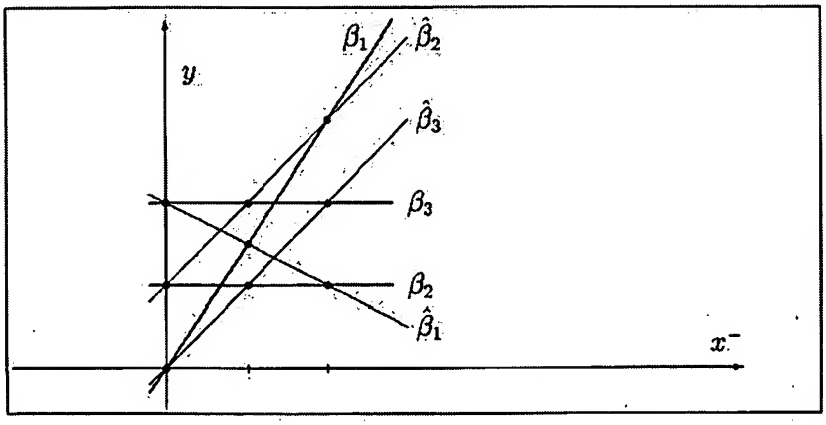


Abbildung 6: Modell 1 nicht identifizierbar

Das Gegenbeispiel 5.4 aus Modell 2 läßt sich durch Verwendung empirischer Verteilungen auf Modell 3 übertragen. Allgemein kann man eine Situation mit festen Regressoren auch in der Terminologie von Modell 3 ausdrücken, wenn man die empirische Verteilung

der Regressoren benutzt. Die folgende Bemerkung und die Bemerkung 6.10 zeigen, wie sich in diesem Fall die Identifikationsaussagen übertragen. Mit etwas höherem Aufwand lassen sich für $p > 1$ aus Beispiel 5.4 auch Beispiele für Modell 3 gewinnen, bei denen sich die Regressorenverteilung auf die entsprechenden $p-1$-dimensionalen Hyperebenen konzentriert, aber keine empirische Verteilung ist. Das werde ich aber hier nicht weiter ausführen.

Die folgende Bemerkung ist destruktiv formuliert („Aus Nichtidentifizierbarkeit in Modell 2 folgt Nichtidentifizierbarkeit in Modell 3"), weil die teilweise Identifizierbarkeitsaussage, die man für Modell 2 direkt erhält (Satz 6.4), stärker ist als das, was sich aus Bemerkung 5.6 gewinnen ließe.

Die umgekehrte Implikation läßt sich mit vertretbarem Aufwand nicht zeigen, da sich eine Verteilung F_J wie in Modell 3 wegen der Mischkomponenten-Anteile ϵ_i im allgemeinen nicht einfach als Fixed-Partition-Verteilung gemäß Modell 2 aufschreiben läßt.

Bemerkung 5.6 *Wenn $\tilde{x}$ so gewählt ist, daß C_{Ω_p} nicht teilweise identifizierbar bzgl. „$\sim_{p1}$" ist, dann ist auch C_{Ω_s} nicht teilweise identifizierbar bzgl. „$\sim_{s1}$", sofern*

$$|I| = n < \infty,$$
$$\mathcal{G} \supset \left\{ \sum_{i \in I} \xi_i \delta_{x_i} : \ \xi_i \geq 0 \quad \forall i \in I, \ \sum_{i \in I} \xi_i = 1 \right\}.$$

Beweis: Sei C_{Ω_p} nicht teilweise identifizierbar bzgl. „$\sim_{p1}$" und

$$\gamma \not\sim_{p1} \hat{\gamma} \tag{5.3}$$
$$\bigotimes_{i \in I} F_{x_i, \gamma(i)} = \bigotimes_{i \in I} F_{x_i, \hat{\gamma}(i)}, \tag{5.4}$$

wobei $F_{x_i,\gamma(i)}$ definiert ist wie in Modell 2. Mit $\gamma(I) = \{(\beta_1, \sigma_1^2), \ldots, (\beta_s, \sigma_s^2)\}$ sei nun weiterhin

$$s := |\gamma(I)|, \quad \forall j = 1, \ldots, s: \ n_j := |\gamma^{-1}\{(\beta_j, \sigma_j^2)\}|,$$
$$G_j := \sum_{\gamma(i) = (\beta_j, \sigma_j^2)} \frac{1}{n_j} \delta_{x_i},$$
$$J \in \Omega_s \text{ definiert durch } S(J) = \{(\beta_j, \sigma_j^2, G_j) : j = 1, \ldots, s\},$$
$$J\{(\beta_j, \sigma_j^2, G_j)\} := \tfrac{n_j}{|I|}, \ j = 1, \ldots, s,$$

und entsprechend $\hat{s}, \hat{\beta}_j, \hat{\sigma}_j^2, \hat{G}_j, \hat{n}_j, \hat{J}$. (5.3) impliziert $S(J) \neq S(\hat{J})$ und damit $J \not\sim_{s1} \hat{J}$. F_J und $F_{\hat{J}}$ seien definiert wie in Modell 3. Es wird nun $F_J = F_{\hat{J}}$ gezeigt, was bedeutet, daß C_{Ω_s} nicht teilweise identifizierbar bzgl. „$\sim_{s1}$" ist. Für Verteilungen F auf $\mathbb{R}^{p+2}$ bezeichne F^X die Verteilung der ersten $p+1$ Komponenten (Regressorenvektor) und $F^{Y|X}$ die bedingte Verteilung der $p+2$. Komponente (abhängige Variable) unter den ersten $p+1$ Komponenten. Ich zeige $F_J^X = F_{\hat{J}}^X$ und $F_J^{Y|X} = F_{\hat{J}}^{Y|X}$. Zunächst gilt $S(F_J^X) = S(F_{\hat{J}}^X) = \{x_i : i \in I\}$ und für alle $i \in I$:

$$F_J^X\{x_i\} = \sum_{j=1}^{s} J\{(\beta_j, \sigma_j^2, G_j)\} G_j\{x_i\} = \frac{|\{k : x_k = x_i\}|}{|I|} = F_{\hat{J}}^X\{x_i\}.$$

Um $F_J^{Y|X} = F_j^{Y|X}$ zu zeigen, wird für $i \in I, B \in \mathcal{B}$ folgende Gleichung benötigt, die aus (5.4) folgt:

$$\sum_{j=1}^{s} \mathcal{N}_{(x_i'\beta_j,\sigma_j^2)}(B)|\{k:\ x_k = x_i, \gamma(k) = (\beta_j,\sigma_j^2)\}| =$$
$$= \sum_{k:x_k=x_i} \mathcal{N}_{(x_i'\beta(k),\sigma^2(k))}(B) = \sum_{k:x_k=x_i} \mathcal{N}_{(x_i'\hat{\beta}(k),\hat{\sigma}^2(k))}(B) =$$
$$= \sum_{j=1}^{s} \mathcal{N}_{(x_i'\hat{\beta}_j,\hat{\sigma}_j^2)}(B)|\{k:\ x_k = x_i, \hat{\gamma}(k) = (\hat{\beta}_j,\hat{\sigma}_j^2)\}|.$$

Daher gilt mit $i \in I$:

$$\forall B \in \mathcal{B}: F_J^{Y|X}(B|x_i) = \frac{F_J\{(x,y):\ y\in B, x=x_i\}}{F_J\{(x,y):\ x=x_i\}} =$$
$$= \frac{\sum_{j=1}^{s} n_j F[\{(x,y):\ y\in B, x=x_i\}, (\beta_j,\sigma_j^2)]}{|I|F_J^X\{x_i\}} =$$
$$= \frac{\sum_{j=1}^{s} n_j \mathcal{N}_{(x_i'\beta_j,\sigma_j^2)}(B) G_j\{x_i\}}{|\{k:\ x_k=x_i\}|} =$$
$$= \frac{\sum_{j=1}^{s} n_j \mathcal{N}_{(x_i'\beta_j,\sigma_j^2)}(B) \frac{|\{k:\ x_k=x_i, \gamma(k)=(\beta_j,\sigma_j^2)\}|}{n_j}}{|\{k:\ x_k=x_i\}|} =$$
$$= \frac{\sum_{j=1}^{s} \mathcal{N}_{(x_i'\beta_j,\hat{\sigma}_j^2)}(B)|\{k:\ x_k=x_i, \hat{\gamma}(k)=(\hat{\beta}_j,\hat{\sigma}_j^2)\}|}{|\{k:\ x_k=x_i\}|} = F_j^{Y|X}(B|x_i),$$

also $F_J = F_j$.

6 Identifizierbarkeitsresultate

In diesem Abschnitt werden hinreichende Bedingungen für Identifizierbarkeit und teilweise Identifizierbarkeit aller drei Modelle gegeben. Dabei ist folgende Voraussetzung immer wesentlich: Die Anzahl der Mischungskomponenten muß kleiner sein als die Anzahl der $p-1$-dimensionalen Hyperebenen, die man benötigt, um die Regressoren, die zu einer einzelnen Mischungskomponente (Cluster) gehören, zu überdecken.

Bemerkung 6.1 *Von der Normalverteilungsvoraussetzung für den Störterm u_i wird hier nur die Identifizierbarkeit der Familie der endlichen univariaten Normal(a,σ^2)-Verteilungsmischungen bzgl. „$\sim_T$" definiert in Bemerkung 4.6 mit $(a,\sigma^2) \in T = \mathbb{R} \times \mathbb{R}_0^+$ benötigt. Ein einfacher Beweis dafür findet sich in Titterington, Smith und Makov (1985), S. 38.*

Außer für Satz 6.7, wo mehrdimensionale Normalverteilungen gebraucht werden, würden daher folgende Voraussetzungen ausreichen: $\mathcal{L}(u_i) = H_\theta$, wobei $E(H_\theta) = 0$, $H_\theta \in \mathcal{F} := \{H_\theta:\ \theta \in T\} \subset \mathcal{P}_1$, $C_{J(T)}$ identifizierbar bzgl. „$\sim_T$",

$$H_\theta(\bullet - t) \in \mathcal{F}\ \forall \theta \in T, t \in \mathbb{R},$$

d.h. Lokationsmischungen der H_θ müssen identifizierbar sein.

Satz 6.2 (Modell 1) *$C_{\mathcal{J}(T_f)}$ ist identifizierbar bzgl. „$\sim_f$" falls an $\mathcal{J}(T_f)$ die zusätzliche Einschränkung $|S(J)| < h$ gemacht wird, wobei*

$$h := \min_q \left\{ \{x_i^- : i \in I\} \subseteq \bigcup_{i=1}^{q} H_i : H_i \in \mathcal{H}_p \right\}.$$

Beweis:

$$F_{x,J} = F_{x,\hat{J}} \Leftarrow J = \hat{J} \quad \forall J, \hat{J} \in \mathcal{J}(T_f)$$

ist klar, es bleibt nur „$\Rightarrow$" (d.h. teilweise Identifizierbarkeit) zu zeigen. Beweis durch Widerspruch: Es sei

$$F_{x,J} = F_{x,\hat{J}}, \; J \neq \hat{J}, \; S(J) := \{(\beta_1, \sigma_1^2), \ldots, (\beta_s, \sigma_s^2)\}$$

und ohne Einschränkung

$$|S(J)| \geq |S(\hat{J})|, \quad J\{(\beta_1, \sigma_1^2)\} \neq \hat{J}\{(\beta_1, \sigma_1^2)\}. \tag{6.1}$$

Angenommen, es gäbe nun $(\beta, \sigma^2) \in S(J)$, so daß

$$\forall i \in I \quad \exists (\hat{\beta}, \hat{\sigma}^2) \in S(\hat{J}) : \; \beta \neq \hat{\beta}, \; x_i'\beta = x_i'\hat{\beta}.$$

Dann folgte

$$\bigcup_{(\hat{\beta},\hat{\sigma}^2) \in S(\hat{J}):\; \hat{\beta} \neq \beta} \{x^- : x'\beta = x'\hat{\beta}\} \supset \{x_i^- : i \in I\} \Rightarrow h \leq |S(\hat{J})|$$

im Widerspruch zu $|S(J)| < h$. Also

$$\forall (\beta, \sigma^2) \in S(J) \quad \exists i(\beta) \in I \quad \forall (\hat{\beta}, \hat{\sigma}^2) \in S(\hat{J}) : \\ x_{i(\beta)}'\beta = x_{i(\beta)}'\hat{\beta} \Rightarrow \beta = \hat{\beta}. \tag{6.2}$$

Sei nun $t := x_{i(\beta_1)}$.
Aus $F_{x,J} = F_{x,\hat{J}}$ folgt

$$F_{t,J} = \int_{T_f} \mathcal{N}_{(t'\beta, \sigma^2)} dJ(\beta, \sigma^2) =$$
$$= F_{t,\hat{J}} = \int_{T_f} \mathcal{N}_{(t'\beta, \sigma^2)} d\hat{J}(\beta, \sigma^2).$$

Daraus folgt, weil endliche Normalverteilungsmischungen identifizierbar sind,

$$(t'\beta_1, \sigma_1^2) \in \{(t'\hat{\beta}, \hat{\sigma}^2) : (\hat{\beta}, \hat{\sigma}^2) \in S(\hat{J})\}$$

und wegen der Definition von t mit (6.2):

$$(\beta_1, \sigma_1^2) \in S(\hat{J}). \tag{6.3}$$

Die Definition von t bringt weiterhin

$$\forall S(\hat{J}) \ni (\hat{\beta}, \hat{\sigma}^2) \neq (\beta_1, \sigma_1^2): \quad (t'\hat{\beta}, \hat{\sigma}^2) \neq (t'\beta_1, \sigma_1^2), \tag{6.4}$$

so daß, durch Identifikation der Normalverteilungsmischung $F_{t,J} = F_{t,\hat{J}}$,

$$\hat{J}\{(\beta_1, \sigma_1^2)\} = \hat{J}\{(\hat{\beta}, \hat{\sigma}^2): (t'\hat{\beta}, \hat{\sigma}^2) = (t'\beta_1, \sigma_1^2)\} = \\ = J\{(\beta, \sigma^2): (t'\beta, \sigma^2) = (t'\beta_1, \sigma_1^2)\}.$$

Daraus folgt $\hat{J}\{(\beta_1, \sigma_1^2)\} > J\{(\beta_1, \sigma_1^2)\}$, denn nach (6.1) ist $\hat{J}\{(\beta_1, \sigma_1^2)\} \neq J\{(\beta_1, \sigma_1^2)\}$. Daher

$$\exists S(J) \ni (\beta_2, \sigma_2^2) \neq (\beta_1, \sigma_1^2): \quad (t'\beta_2, \sigma_2^2) = (t'\beta_1, \sigma_1^2). \tag{6.5}$$

Anwendung derselben Argumentation auf $x_{i(\beta_2)}$ bringt analog zu (6.3):

$$(\beta_2, \sigma_2^2) \in S(\hat{J}).$$

Das ist ein Widerspruch zu (6.4) und (6.5). Also folgt $J = \hat{J}$.

Korollar 6.3 (Modell 1, keine Einschränkungen für J) *$C_{\mathcal{J}(T_f)}$ ist identifizierbar bzgl. „$\sim_f$" falls*

$$\forall m \in \mathbb{N}, A = \bigcup_{i=1}^{m} H_i: \{x_i^- : i \in I\} \not\subset A \tag{6.6}$$

für beliebige $p-1$-dimensionale Hyperebenen $H_i \in \mathcal{H}_p, i = 1, \ldots, m$.

Beweis: (6.6) $\Leftrightarrow \infty = h$ (definiert in Satz 6.2.)

Satz 6.4 (Modell 2) *C_{Ω_p} ist teilweise identifizierbar bzgl. „$\sim_{pl}$" falls an Ω_p folgende zusätzliche Einschränkung gemacht wird: $|\gamma(I)| < \min_{(\beta,\sigma^2)\in\gamma(I)} h(\beta, \sigma^2)$, wobei*

$$h(\beta, \sigma^2) := \min_q \left\{ \{x_i^- : i \in I, \gamma(i) = (\beta, \sigma^2)\} \subseteq \bigcup_{j=1}^{q} H_j : H_j \in \mathcal{H}_p \ \forall j \right\}.$$

Beweis: Angenommen, C_{Ω_p} sei nicht teilweise identifizierbar bzgl. „$\sim_{pl}$". Dann gibt es

$$\gamma \not\sim_{pl} \hat{\gamma} \in \Omega \text{ mit } F_{x,\gamma} = F_{x,\hat{\gamma}}, \\ \text{wobei ohne Einschränkung } |\gamma(I)| \geq |\hat{\gamma}(I)|, \quad (\beta_0, \sigma_0^2) \notin \hat{\gamma}(I). \tag{6.7}$$

Ich beweise weiter unten

$$\forall (\hat{\beta}, \hat{\sigma}^2) \in \hat{\gamma}(I): \quad \dim(X_{(\hat{\beta},\hat{\sigma}^2)} \times \{1\}) < p+1, \tag{6.8}$$

wobei $X_{(\hat{\beta},\hat{\sigma}^2)} := \{x_i^- : i \in I, \gamma(i) = (\beta_0, \sigma_0^2), \hat{\gamma}(i) = (\hat{\beta}, \hat{\sigma}^2)\}$. Das heißt:

$$\forall (\hat{\beta}, \hat{\sigma}^2) \in \hat{\gamma}(I) \quad \exists H_{(\hat{\beta},\hat{\sigma}^2)} \in \mathcal{H}_p: \quad X_{(\hat{\beta},\hat{\sigma}^2)} \subset H_{(\hat{\beta},\hat{\sigma}^2)}.$$

Daraus folgt

$$h(\beta_0, \sigma_0^2) \leq |\hat{\gamma}(I)| \leq |\gamma(I)|$$

im Widerspruch zu $|\gamma(I)| < \min\{h(\beta, \sigma^2)\}$. Daher

$$F_{x,\gamma} = F_{x,\hat{\gamma}} \Rightarrow \gamma \sim_{p1} \hat{\gamma},$$

also teilweise Identifizierbarkeit von C_{Ω_p} bzgl. „$\sim_{p1}$".
Beweis von (6.8): Es gilt

$$\bar{x_i} \in X_{(\hat{\beta},\hat{\sigma}^2)} \Rightarrow F_{\bar{x}_i,\gamma(i)} = F_{\bar{x}_i,\hat{\gamma}(i)} \Rightarrow \mathcal{N}_{(x_i'\beta_0,\sigma_0^2)} = \mathcal{N}_{(x_i'\hat{\beta},\hat{\sigma}^2)}.$$

Daraus ergibt sich

$$\forall \bar{x_i} \in X_{(\hat{\beta},\hat{\sigma}^2)}: \; x_i'\beta_0 = x_i'\hat{\beta}, \; \sigma_0^2 = \hat{\sigma}^2, \text{ also}$$
$$(\beta_0, \sigma_0^2) = (\hat{\beta}, \hat{\sigma}^2) \text{ oder } \dim\langle X_{(\hat{\beta},\hat{\sigma}^2)} \times \{1\}\rangle < p+1,$$

wobei ersteres in (6.7) ausgeschlossen wurde.

Korollar 6.5 (Modell 2, keine Einschränkungen für $|\gamma(I)|$)
C_{Ω_p} ist teilweise identifizierbar bzgl. „$\sim_{p1}$" unter der zusätzlichen Einschränkung an Ω_p:

$$\forall m \in \mathbb{N}, \; (\beta, \sigma^2) \in \gamma(I), \; A = \bigcup_{i=1}^{m} H_i:$$
$$\{\bar{x_i} : i \in I, \gamma(i) = (\beta, \sigma^2)\} \not\subset A \tag{6.9}$$

für beliebige $p-1$-dimensionale Hyperebenen $H_i \in \mathcal{H}_p, i = 1, \ldots, m$.

Beweis: (6.9) $\Leftrightarrow \infty = \min_{(\beta,\sigma^2)\in\gamma(I)} h(\beta, \sigma^2)$ (definiert in Satz 6.4).

Bemerkung 6.6 *In Satz 6.2 und Korollar 6.3 kann h anhand des theoretischen Designs $\{x_i : i \in I\}$ berechnet werden. Man braucht also nichts über die unbekannten Modellparameter zu wissen, um eine Identifizierbarkeitsaussage zu machen. Das ist in Satz 6.4 und Korollar 6.5 nicht möglich, denn $h(\beta, \sigma^2)$ hängt von Ω_p durch die unbekannte Zuordnung der Punkte zu den Clustern ab.*

Satz 6.7 (Modell 3: Normalverteilte Regressoren) *$C_{\mathcal{J}(T_s)}$ ist unter der zusätzlichen Einschränkung $\sigma^2 > 0$ an T_s identifizierbar bzgl. „$\sim_{T_s}$", falls*

$$\mathcal{G} = \{\mathcal{N}_{(a,\Sigma)} : a \in \mathbb{R}^p, \Sigma \in \mathcal{S}_p\},$$

wobei $\mathcal{S}_p$ die Menge der symmetrischen positiv definiten reellwertigen $p \times p$-Matritzen bezeichne.

Beweis: Es wird unten gezeigt, daß es eine bijektive Beziehung zwischen den Parametern einer $p+1$-variaten Normalverteilung und $\theta \in T_s$ aus Modell 3 gibt. Damit ist Modell 3 eine Umparametrisierung einer endlichen Mischung von $p+1$-variaten Normalverteilungen. Solche Mischungen sind identifizierbar (Yakowitz and Spragins (1968)).

G ist festgelegt durch a und Σ, $F(\bullet, \theta)$ also durch

$$T(\theta) := \left(\beta, \sigma^2, a, (s_{ij})_{i,j=1,\dots,p}\right) \in I\!R^{p+1} \times I\!R^+ \times I\!R \times \mathcal{S}_p,$$

wobei $(s_{ij})_{i,j=1,\dots,p} := \Sigma$. Gesucht ist also eine bijektive Abbildung $I\!R^{p+1} \times I\!R^+ \times I\!R \times \mathcal{S}_p \mapsto I\!R^{p+1} \times \mathcal{S}_{p+1}$, die $T(\theta)$ auf (b, Γ) abbildet, so daß $F(\bullet, \theta) = \Phi_{b,\Gamma}$.

$F(\bullet, \theta)$ ist als Verteilung einer linearen Funktion der normalverteilten Zufallsvariablen x^- und u eine $p+1$-variate Normalverteilung. Sei also $T(\theta)$ wie oben gegeben, $\alpha := \beta^-$, also $\beta = (\alpha', \beta_{p+1})'$. Die ersten p Komponenten sind die Parameter für die Regressoren $(x_1, \dots, x_p)$. $(b, \Gamma) =: (b, (t_{ij})_{i,j=1,\dots,p+1})$ ergibt sich dann wie folgt:

$$\begin{gathered}
(b_1, \dots, b_p)' = Ex = a, \quad b_{p+1} = Ey = \alpha' a + \beta_{p+1}, \\
(t_{ij})_{i,j=1,\dots,p} = \operatorname{Cov} x = \Sigma, \\
i = 1, \dots, p: \; t_{p+1,i} = t_{i,p+1} = \operatorname{Cov} x_i y = \textstyle\sum_{j=1}^p \beta_j s_{ij}, \\
t_{p+1,p+1} = \operatorname{Var} y = \alpha' \Sigma \alpha + \sigma^2.
\end{gathered}$$

Diese Abbildung ist bijektiv, da sich bei gegebenem (b, Γ) aus den obigen Gleichungen folgendermaßen die Umkehrabbildung definieren läßt (wieder sei $\alpha = \beta^-$):

$$a = (b_1, \dots, b_p), \quad \Sigma = (t_{ij})_{i,j=1,\dots,p}, \tag{6.10}$$

$$\begin{pmatrix} a & 1 \\ \Sigma & 0 \end{pmatrix} \beta = \begin{pmatrix} b_{p+1} \\ t_{1,p+1} \\ \dots \\ t_{p,p+1} \end{pmatrix}, \tag{6.11}$$

$$\sigma^2 = t_{p+1,p+1} - \alpha' \Sigma \alpha.$$

Dabei ist β eindeutig definiert, da $\begin{pmatrix} a & 1 \\ \Sigma & 0 \end{pmatrix}$ mit Σ invertierbar ist, und es ist $\sigma^2 > 0$, da für $\mathcal{L}(x', y)' = \mathcal{N}_{(b,\Gamma)}$ mit den obigen Bezeichnungen

$$\begin{aligned}
0 < \operatorname{Var}(y - \alpha' x) &= \operatorname{Var} y + \operatorname{Var}(\alpha' x) - 2 \operatorname{Cov}(y, \alpha' x) = \\
&= t_{p+1,p+1} + \alpha' \Sigma \alpha - 2\alpha' \begin{pmatrix} t_{1,p+1} \\ \dots \\ t_{p,p+1} \end{pmatrix} = t_{p+1,p+1} - \alpha' \Sigma \alpha = \sigma^2,
\end{aligned}$$

denn wegen $\alpha = \beta^-$ und (6.11) ist

$$\operatorname{Var}(\alpha' x) = \alpha' \Sigma \alpha = \alpha' \begin{pmatrix} t_{1,p+1} \\ \dots \\ t_{p,p+1} \end{pmatrix}.$$

(Man beachte, daß $\mathcal{G}$ keine endlichen Mischungen seiner Elemente mit mehr als einer Komponente enthält. Daher taucht das Problem aus Beispiel 5.1 nicht auf.)

Satz 6.8 (Modell 3: Keine Masse auf $p-1$-dimensionalen Hyperebenen)
C_{Ω_s} ist identifizierbar bzgl. „$\sim_s$" falls

$$\mathcal{G} \subseteq \{P \in \mathcal{P}_p : P(H) = 0 \quad \forall H \in \mathcal{H}_p\}.$$

Beweis: Der Beweis benötigt folgendes Resultat: Wenn $P^{(X,Y)} = P^X \otimes P^{(Y|X)}$, $P^{(X,Z)}$ analog, $P^{(Y|X=x)}$ und $P^{(Z|X=x)}$ definiert auf σ-Algebren $\mathcal{B}$, wobei

$$\exists \mathcal{E} \text{ abzählbar: } \sigma(\mathcal{E}) = \mathcal{B},$$

dann

$$P^{(X,Y)} = P^{(X,Z)} \Leftrightarrow \exists A \subseteq \left\{x : P^{(Y|X=x)} = P^{(Z|X=x)}\right\} : P^X(A) = 1. \tag{6.12}$$

Beweis: Gänssler / Stute (1977), S. 197.

Nun seien $s := |S(J)|$, $\hat{s} := |S(\hat{J})|$,

$$S(J) =: \{(\beta_i, \sigma_i^2, G_i), i = 1, \ldots, s\}, \quad \epsilon_i := J\{(\beta_i, \sigma_i^2, G_i)\},$$
$$S(\hat{J}) =: \{(\hat{\beta}_i, \hat{\sigma}_i^2, G_i), i = 1, \ldots, \hat{s}\}, \quad \hat{\epsilon}_i := \hat{J}\{(\hat{\beta}_i, \hat{\sigma}_i^2, \hat{G}_i)\}.$$

Mit diesen Bezeichnungen ist

$$F_J = \sum_{i=1}^{s} \epsilon_i F(\bullet, \beta_i, \sigma_i^2, G_i),$$

F_J wie in Modell 3, entsprechend $F_{\hat{J}}$. Natürlich gilt wieder

$$F_J = F_{\hat{J}} \Leftarrow J = \hat{J} \quad \forall J, \hat{J} \in \Omega_s,$$

so daß nur

$$F_J = F_{\hat{J}} \Rightarrow J = \hat{J} \tag{6.13}$$

(d.h. teilweise Identifizierbarkeit) zu zeigen ist.

Übersicht: Die bedingten Verteilungen für y bei gegebenem x sind eindimensionale Normalverteilungen und damit identifizierbar. Insbesondere sind also die $(x'\beta_i, \sigma_i^2)$ identifizierbar; damit auch durch Wahl einer geeigneten Menge M von Regressoren die (β_i, σ_i^2), also $S(J) = S(\hat{J})$. Schließlich identifiziert man die ϵ_i durch Integration der Anteile der Mischungskomponenten im bedingten Fall über x und kann damit (nach entsprechender Umnumerierung) auch noch $G_i = \hat{G}_i$, $i = 1, \ldots, s$, zeigen, also $J = \hat{J}$.

Vorbereitung: Definiere

$$\mu := \sum_{i=1}^{s} G_i + \sum_{i=1}^{\hat{s}} \hat{G}_i,$$
$$F^{(X,Y)} := F_J, \; F^{(\hat{X},\hat{Y})} := F_{\hat{J}}, \; g_i(x^-) := \frac{dG_i}{d\mu}(x^-), \quad \hat{g}_i(x^-) := \frac{d\hat{G}_i}{d\mu}(x^-),$$

wobei ich in diesem Beweis ausnahmsweise die Zufallsvariablen der Regressoren $X, \hat{X}$ ($\mathbb{R}^p$-wertig, d.h. die $p+1$. Komponente 1 wird nicht als Bestandteil der Zufallsvariable interpretiert) und abhängigen Variablen Y bzw. $\hat{Y}$ mit Großbuchstaben bezeichne und die von ihnen angenommenen Werte mit Kleinbuchstaben, wobei aber wie gewohnt $x \in \mathbb{R}^p \times \{1\}$ sei. Damit

$$F^{Y|X=x^-} = \sum_{i=1}^{s} \epsilon_i \frac{g_i(x^-)}{\sum_{j=1}^{s} \epsilon_j g_j(x^-)} \mathcal{N}_{(x'\beta_i, \sigma_i^2)}, \text{ da} \tag{6.14}$$

$$\forall B_1 \in \mathcal{B}^p, B_2 \in \mathcal{B}: \; F^{(X,Y)}(B_1 \times B_2) = \sum_{i=1}^{s} \epsilon_i \int_{B_1} \mathcal{N}_{(x'\beta_i, \sigma_i^2)}(B_2) dG_i(x^-) =$$

$$= \int_{B_1} \sum_{i=1}^{s} \epsilon_i g_i(x^-) \mathcal{N}_{(x'\beta_i, \sigma_i^2)}(B_2) d\mu(x^-) =$$

$$= \int_{B_1} \sum_{i=1}^{s} \epsilon_i \frac{g_i(x^-)}{\sum_{j=1}^{s} \epsilon_j g_j(x^-)} \mathcal{N}_{(x'\beta_i, \sigma_i^2)}(B_2) \sum_{j=1}^{s} \epsilon_j g_j(x^-) d\mu(x^-) =$$

$$= \int_{B_1} \sum_{i=1}^{s} \epsilon_i \frac{g_i(x^-)}{\sum_{j=1}^{s} \epsilon_j g_j(x^-)} \mathcal{N}_{(x'\beta_i, \sigma_i^2)}(B_2) d\left(\sum_{j=1}^{s} \epsilon_j G_j\right)(x^-) =$$

$$= \int_{B_1} F^{Y|X=x^-}(B_2) dF^X(x^-).$$

Analog zu (6.14) berechnet man

$$F^{\hat{Y}|\hat{X}=x^-} = \sum_{i=1}^{\hat{s}} \hat{\epsilon}_i \frac{\hat{g}_i(x^-)}{\sum_{j=1}^{\hat{s}} \hat{\epsilon}_j \hat{g}_j(x^-)} \mathcal{N}_{(x'\hat{\beta}_i, \hat{\sigma}_i^2)}. \tag{6.15}$$

Sei nun

$$M := \{x^- : \quad \forall j,k \in \{1,\ldots,s\}, l,m \in \{1,\ldots,\hat{s}\}:$$
$$x'\beta_j = x'\beta_k \Rightarrow \beta_j = \beta_k, \quad x'\beta_j = x'\hat{\beta}_l \Rightarrow \beta_j = \hat{\beta}_l,$$
$$x'\hat{\beta}_l = x'\hat{\beta}_m \Rightarrow \hat{\beta}_l = \hat{\beta}_m\} =$$

$$= \mathbb{R}^p \setminus \left[\bigcup_{S(J) \ni \beta_j \neq \beta_k \in S(J)} \{x^- : x'\beta_j = x'\beta_k\} \cup \right.$$

$$\left. \cup \bigcup_{S(J) \ni \beta_j \neq \hat{\beta}_l \in S(\hat{J})} \{x^- : x'\beta_j = x'\hat{\beta}_l\} \cup \bigcup_{S(\hat{J}) \ni \hat{\beta}_l \neq \hat{\beta}_m \in S(\hat{J})} \{x^- : x'\hat{\beta}_l = x'\hat{\beta}_m\} \right].$$

M ist also Komplement einer endlichen Vereinigung von Elementen aus $\mathcal{H}_p$. Daher folgt aus der Voraussetzung an $\mathcal{G}$:

$$F^X = \sum_{i=1}^{s} \epsilon_i G_i \Rightarrow F^X(M) = 1.$$

Für $x^- \in M$ sind alle $(x'\beta_i, \sigma_i^2)$, $i = 1, \ldots, s$, paarweise verschieden, da alle (β_i, σ_i^2), $i = 1, \ldots, s$ wegen $J \in \Omega_s$ aus Beispiel 4.11 paarweise verschieden sind.

Identifikation der (β_i, σ_i^2): Sei nun für $x^- \in M$ eine empirische Verteilung J_x auf $\mathbb{R} \times \mathbb{R}_0^+$ definiert durch

$$J_x\{(x'\beta_i, \sigma_i^2)\} := \epsilon_i \frac{g_i(x^-)}{\sum_{j=1}^{s} \epsilon_j g_j(x^-)}, \quad i = 1, \ldots, s, \tag{6.16}$$
$$S(J_x) = \{(x'\beta_i, \sigma_i^2) : i = 1, \ldots, s; \quad g_i(x^-) > 0\},$$

so daß mit (6.14)

$$F^{Y|X=x^-} = \int_{\mathbb{R}\times\mathbb{R}_0^+} \mathcal{N}_\theta dJ_x(\theta).$$

und sei $\hat{J}_x$ definiert analog zu (6.16), so daß mit (6.15)

$$F^{\hat{Y}|\hat{X}=x^-} = \int_{\mathbb{R}\times\mathbb{R}_0^+} \mathcal{N}_\theta d\hat{J}_x(\theta).$$

Sei nun $F_J = F_{\hat{J}}$, also insbesondere $F^X = F^{\hat{X}}$. Dann impliziert (6.12)

$$\exists N \subseteq M, \quad F^{\hat{X}}(N) = F^X(N) = 1: \quad x^- \in N \Rightarrow F^{Y|X=x^-} = F^{\hat{Y}|\hat{X}=x^-},$$
$$\text{so daß } J_x = \hat{J}_x, \tag{6.17}$$

da endliche Normalverteilungsmischungen identifizierbar bzgl. „$\sim_T$" nach Bemerkung 6.1 sind. Weiterhin:

$$\forall i \in \{1, \ldots, s\}, j \in \{1, \ldots, \hat{s}\}: \quad F^X(N) = 1 \Rightarrow G_i(N) = \hat{G}_j(N) = 1 \Rightarrow$$
$$\Rightarrow \exists x(i) \in N, \hat{x}(j) \in N: \quad g_i(x(i)^-) > 0, \quad \hat{g}_j(\hat{x}(j)^-) > 0. \tag{6.18}$$

Für gegebenes $i \in \{1, \ldots, s\}$ folgt aus (6.17)

$$\exists j \in \{1, \ldots, \hat{s}\}: x(i)'\beta_i = x(i)'\hat{\beta}_j, \quad \sigma_i^2 = \hat{\sigma}_j^2.$$

Aus der Definition von M ergibt sich dann

$$(\beta_i, \sigma_i^2) \in \{(\hat{\beta}_j, \hat{\sigma}_j^2) : j \in \{1, \ldots, \hat{s}\}\}.$$

Dasselbe Argument kann auf $\hat{x}(j) \quad \forall j \in \{1, \ldots, \hat{s}\}$ angewendet werden, so daß

$$(\hat{\beta}_j, \hat{\sigma}_j^2) \in \{(\beta_i, \sigma_i^2) : i \in \{1, \ldots, s\}\}.$$

Zusammen:

$$\{(\beta_i, \sigma_i^2) : i \in \{1, \ldots, s\}\} = \{(\hat{\beta}_j, \hat{\sigma}_j^2) : j \in \{1, \ldots, \hat{s}\}\}.$$

Aufgrund der Definition von Ω_s aus Beispiel 4.11 sind sowohl die $(\beta_i, \sigma_i^2) : i \in \{1, \ldots, s\}$ als auch die $(\hat{\beta}_j, \hat{\sigma}_j^2) : j \in \{1, \ldots, \hat{s}\}$ paarweise verschieden, so daß sich nun $s = \hat{s}$ ergibt und man ohne Einschränkung $(\beta_i, \sigma_i^2) = (\hat{\beta}_i, \hat{\sigma}_i^2)$, $i = 1, \ldots, s$ annehmen kann.

Identifikation von G_i, ϵ_i: Weiterhin gibt es für $i = 1, \ldots, s$ eindeutige G_i und $\epsilon_i = J\{(\beta_i, \sigma_i^2, G_i)\}$, $\hat{G}_i$ und $\hat{\epsilon}_i = \hat{J}\{(\beta_i, \sigma_i^2, \hat{G}_i)\}$ mit

$$(\beta_i, \sigma_i^2, G_i) \in S(J), \quad (\beta_i, \sigma_i^2, \hat{G}_i) \in S(\hat{J}).$$

Definiere für $x \in N$, $i = 1, \ldots, s$:

$$\xi_i(x) := J_x\{(x'\beta_i, \sigma_i^2)\}, \quad \hat{\xi}_i(x) := \hat{J}_x\{(x'\beta_i, \sigma_i^2)\}.$$

Aus (6.17) ergibt sich dann

$$\forall x \in N : \xi_i(x) = \hat{\xi}_i(x).$$

Weil $F^X(N) = 1$ und daher $\forall i = 1, \ldots, s : \int_N g_i(x^-)d\mu(x^-) = 1$, erhalten wir mit (6.16)

$$\epsilon_i = \int_N \epsilon_i g_i(x^-)d\mu(x^-) = \int_N \xi_i(x) \sum_{j=1}^{s} \epsilon_j g_j(x^-)d\mu(x^-) =$$

$$= \int_N \xi_i(x)dF^X(x^-) = \int_N \hat{\xi}_i(x)dF^X(x^-) = \hat{\epsilon}_i.$$

$$\text{Zuletzt, für } x \in N : g_i(x^-) = \frac{\xi_i(x)}{\epsilon_i} \sum_{j=1}^{s} \epsilon_j g_j(x^-) =$$

$$= \frac{\hat{\xi}_i(x)}{\hat{\epsilon}_i} \frac{dF^X}{d\mu}(x^-) = \hat{g}_i(x^-) \Rightarrow G_i = \hat{G}_i \Rightarrow$$

$$\Rightarrow J = \hat{J}.$$

Satz 6.9 (Teilweise Identifizierbarkeit von Modell 3) *C_{Ω_s} ist teilweise identifizierbar bzgl. „$\sim_{s1}$" falls*

$$\mathcal{G} \subset \mathcal{P}_p \text{ mit } \forall G \in \mathcal{G}, m \in \mathbb{N}, A = \bigcup_{i=1}^{m} H_i : \quad G(A) < 1 \tag{6.19}$$

für beliebige $p-1$-dimensionale Hyperebenen $H_i \in \mathcal{H}_p, i = 1, \ldots, m$.

Beweis: Es gelten die Bezeichnungen aus dem Beweis von Satz 6.8. (6.14) und (6.15) benötigen keine Voraussetzung an $\mathcal{G}$ und gelten daher auch hier. Es wird analog zum Beweis von Satz 6.8 vorgegangen.

Sei wieder

$$M := \{x^- : \forall j, k \in \{1, \ldots, s\}, l, m \in \{1, \ldots, \hat{s}\} :$$
$$x'\beta_j = x'\beta_k \Rightarrow \beta_j = \beta_k, \quad x'\beta_j = x'\hat{\beta}_l \Rightarrow \beta_j = \hat{\beta}_l,$$
$$x'\hat{\beta}_l = x'\hat{\beta}_m \Rightarrow \hat{\beta}_l = \hat{\beta}_m\}.$$

F^X ist eine endliche Konvexkombination von Elementen von $\mathcal{G}$ und erfüllt daher (6.19). Für M, das Komplement einer endlichen Vereinigung von Elementen von $\mathcal{H}_p$, gilt daher $F^X(M) > 0$.

Für $x^- \in M$ sind alle $(x'\beta_i, \sigma_i^2)$, $i = 1, \ldots, s$, paarweise verschieden, da alle (β_i, σ_i^2), $i = 1, \ldots, s$ wegen $J \in \Omega_s$ aus Beispiel 4.11 paarweise verschieden sind. Für $x \in M$ seien nun J_x bzw. $\hat{J}_x$ wieder gemäß (6.16) definiert, also mit (6.14)

$$F^{Y|X=x} = \int_{\mathbb{R}\times\mathbb{R}_0^+} \mathcal{N}_\theta dJ_x(\theta),$$

$$\text{sowie } F^{\hat{Y}|\hat{X}=x} = \int_{\mathbb{R}\times\mathbb{R}_0^+} \mathcal{N}_\theta d\hat{J}_x(\theta) \text{ mit (6.15).}$$

Sei nun $F_J = F_{\hat{J}}$, also insbesondere $F^X = F^{\hat{X}}$. Dann impliziert (6.12)

$$\exists N \subseteq M,\quad F^{\tilde{X}}(N) = F^X(N) = F^X(M) > 0: \quad x^- \in N \Rightarrow F^{Y|X=x^-} = F^{\hat{Y}|\hat{X}=x^-},$$
$$\text{so daß } J_x = \hat{J}_x, \tag{6.20}$$

da endliche Normalverteilungsmischungen identifizierbar bzgl. „$\sim_T$" nach Bemerkung 6.1 sind. Weiterhin:

$$\forall i \in \{1,\ldots,s\}, j \in \{1,\ldots,\hat{s}\}: \; G_i, \hat{G}_j \in \mathcal{G} \Rightarrow$$
$$\Rightarrow 0 < G_i(M) = G_i(N), \quad 0 < \hat{G}_j(M) = \hat{G}_j(N) \Rightarrow$$
$$\Rightarrow \exists x(i) \in N, \hat{x}(j) \in N: \quad g_i(x(i)^-) > 0, \quad \hat{g}_j(\hat{x}(j)^-) > 0.$$

Für gegebenes $i \in \{1,\ldots,s\}$ folgt aus (6.20)

$$\exists j \in \{1,\ldots,\hat{s}\}: x(i)'\beta_i = \hat{x}(i)'\hat{\beta}_j, \quad \sigma_i^2 = \hat{\sigma}_j^2.$$

Mit der Definition von M ergibt sich

$$(\beta_i, \sigma_i^2) \in \{(\hat{\beta}_j, \hat{\sigma}_j^2): j \in \{1,\ldots,\hat{s}\}\}.$$

Dasselbe Argument kann auf $\hat{x}(j) \quad \forall j \in \{1,\ldots,\hat{s}\}$ angewendet werden, so daß

$$(\hat{\beta}_j, \hat{\sigma}_j^2) \in \{(\beta_i, \sigma_i^2): i \in \{1,\ldots,s\}\}.$$

Zusammen:

$$\{(\beta_i, \sigma_i^2): i \in \{1,\ldots,s\}\} = \{(\hat{\beta}_j, \hat{\sigma}_j^2): j \in \{1,\ldots,\hat{s}\}\},$$

was bereits $J \sim_{s1} \hat{J}$ bedeutet.

Bemerkung 6.10 *Auch zwischen Modell 1 und Modell 3 besteht ein Zusammenhang bei Verwendung empirischer Verteilungen, so daß sich aus Satz 6.2 bei endlichem I ein Identifizierbarkeitsresultat für Modell 3 gewinnen läßt:*

Identifizierbarkeit von $C_{\mathcal{J}(T_f)}$ bzgl. „$\sim_f$" aus Beispiel 4.9 ist äquivalent zur Identifizierbarkeit von $C_{\mathcal{J}(T_s)}$ bzgl. „$\sim_{T_s}$" aus Beispiel 4.11, falls

$$\mathcal{G} = \left\{G: G = \sum_{i \in I} \xi_i \delta_{x_i}\right\}$$

für beliebige feste $\xi_i > 0$ mit $\sum_{i \in I} \xi_i = 1$. $\mathcal{G}$ enthält hier also nur ein Element, $\Omega_s = \mathcal{J}(T_s)$ und „$\sim_s$"=„$\sim_{T_s}$".

Beweis: Sei zuerst C_{Ω_s} identifizierbar bzgl. „$\sim_s$". Definiere $J, \hat{J} \in \mathcal{J}(T_f)$, $K, \hat{K} \in \Omega_s$ für gegebene (β_i, σ_i^2), $i = 1,\ldots,s$, $(\hat{\beta}_i, \hat{\sigma}_i^2)$, $i = 1,\ldots,\hat{s}$ gemäß

$$J := \sum_{i=1}^{s} \epsilon_i \delta_{(\beta_i, \sigma_i^2)}, \quad \hat{J} := \sum_{i=1}^{\hat{s}} \hat{\epsilon}_i \delta_{(\hat{\beta}_i, \hat{\sigma}_i^2)},$$
$$K := \sum_{i=1}^{s} \epsilon_i \delta_{(\beta_i, \sigma_i^2, G)}, \quad \hat{K} := \sum_{i=1}^{\hat{s}} \hat{\epsilon}_i \delta_{(\hat{\beta}_i, \hat{\sigma}_i^2, G)}.$$

Offenbar gilt

$$J = \hat{J} \Leftrightarrow K = \hat{K}. \tag{6.21}$$

Sei nun $F_J = F_{\hat{J}}$ vorausgesetzt. Das soll äquivalent zu $J = \hat{J}$ sein. Es gilt

$$\begin{aligned}
F_J = F_{\hat{J}} &\Leftrightarrow \bigotimes_{i\in I}\left(\int_{T_I} \mathcal{N}_{(x_i'\beta,\sigma^2)} dJ(\beta,\sigma^2)\right) = \bigotimes_{i\in I}\left(\int_{T_I} \mathcal{N}_{(x_i'\beta,\sigma^2)} d\hat{J}(\beta,\sigma^2)\right) \Leftrightarrow \\
&\Leftrightarrow \forall x_i, i \in I : \int_{T_I} \mathcal{N}_{(x_i'\beta,\sigma^2)} dJ(\beta,\sigma^2) = \int_{T_I} \mathcal{N}_{(x_i'\beta,\sigma^2)} d\hat{J}(\beta,\sigma^2) \Leftrightarrow \\
&\Leftrightarrow G\left\{x : \int_{T_I} \mathcal{N}_{(x'\beta,\sigma^2)} dJ(\beta,\sigma^2) = \int_{T_I} \mathcal{N}_{(x'\beta,\sigma^2)} d\hat{J}(\beta,\sigma^2)\right\} = 1 \Leftrightarrow \\
&\int_{T_s} F(x,y,\beta,\sigma^2,G) dK(\beta,\sigma^2,G) = \int_{T_s} F(x,y,\beta,\sigma^2,G) d\hat{K}(\beta,\sigma^2,G) \Leftrightarrow \\
&\Leftrightarrow F_K = F_{\hat{K}}.
\end{aligned} \tag{6.22}$$

Daraus folgt $K = \hat{K}$ aufgrund der Identifizierbarkeit von C_{Ω_s}. Mit (6.21) ergibt sich nun die Identifizierbarkeit von $C_{\mathcal{J}(T_I)}$ bzgl. „$\sim_I$".

Sei nun umgekehrt $C_{\mathcal{J}(T_I)}$ identifizierbar bzgl. „$\sim_I$" und $K = \hat{K}$ vorausgesetzt. Das ist äquivalent zu $J = \hat{J}$ und weiter wegen der Identifizierbarkeit von $C_{\mathcal{J}(T_I)}$ zu $F_J = F_{\hat{J}}$. Von da an gilt wieder die Äquivalenzumformung (6.22), also insgesamt $F_K = F_{\hat{K}} \Leftrightarrow K = \hat{K}$ und Identifizierbarkeit von C_{Ω_s} bzgl. „$\sim_s$".

Teil II
Fixpunktcluster

7 Einführung: Fixpunktcluster

7.1 Cluster und Ausreißer: Die allgemeine Fixpunktcluster-Idee

Die Idee der Fixpunktcluster wurde bereits in Abschnitt 3.5.1 kurz angedeutet. In diesem Abschnitt werde ich sie allgemein, d.h. nicht an das Regressions-Problem gebunden, erläutern. Anschaulich besteht ein Cluster von Daten in einem Datensatz aus Punkten, die in irgendeiner Weise zusammengehören, während die anderen Punkte nicht dazugehören. „Zusammengehören" heißt meistens „nahe beieinanderliegen". Das ist aber zum Beispiel im Regressionsfall nicht unbedingt so, wie in der Einleitung schon angesprochen wurde. Ein Fixpunktcluster (FPC) soll eine Menge von Punkten sein, so daß die anderen Punkte des Datensatzes bezogen auf die Punkte des FPC Ausreißer sind, von den Punkten des FPC jedoch keiner. In diesem Sinne gehören die Punkte des FPC zusammen, die anderen Punkte gehören nicht dazu. Dieses Konzept wird im folgenden präzisiert.

Vor dem Hintergrund statistischer Modellbildung ist ein Ausreißer bezogen auf eine Verteilung ein Punkt, der nicht zur Verteilung paßt. Zum Beispiel könnte ein Datensatz modelliert werden durch eine Verteilung

$$H = (1-\epsilon)H_0 + \epsilon H^*, \quad \frac{1}{2} > \epsilon \geq 0, \quad H^* \neq H_0. \tag{7.1}$$

In der Literatur heißt ein solches Modell häufig „contamination model" (Verunreinigungsmodell). Die durch H^* erzeugten Punkte wären dann Ausreißer bezüglich H_0. In solchen Situationen kann zum Beispiel getestet werden, ob einzelne Punkte eines Datensatzes von H^* erzeugt wurden. Dieser und andere Ansätze zur Ausreißererkennung werden zum Beispiel in Barnett und Lewis (1984) diskutiert. Solche Verfahren benötigen aber immer eine genauere Spezifikation von H^*. Wie mit Punkten verfahren wird, die weder so aussehen, als seien sie von H_0, noch von H^* erzeugt, ist unklar. Allgemein ist eine Ausreißeridentifikation nach diesem Ansatz nur möglich, wenn H^* Punkte generiert, die sich mit großer Wahrscheinlichkeit deutlich von durch H_0 erzeugten Punkten unterscheiden.

Modelle der Form (7.1) sind auch in der robusten Statistik gebräuchlich. Huber (1981) leitet zum Beispiel Minimax-Schätzer für die Parameter von H_0 her. Diese Schätzer sollen für gegebenes ϵ im „schlimmsten Fall" von H^* die kleinstmögliche Verzerrung bzw. Varianz haben. Häufig liefern robuste Schätzverfahren nebenbei eine Ausreißerklassifikation, siehe dazu Abschnitt 3.5.1.

Davies und Gather (1989) machen einen alternativen Ansatz zur Beschreibung von Ausreißern. Sie beschäftigen sich nur mit dem Fall des eindimensionalen Lokationsproblems mit stetigen Verteilungen[9]. Ich formuliere hier eine Interpretation ihres Ansatzes für allgemeinere Verteilungsklassen.

[9]Die veröffentlichte und überarbeitete Version ihres Papiers (Davies und Gather (1993)) beschränkt sich sogar auf eindimensionale Normalverteilungen.

Ein Punkt wird bei Davies und Gather „Ausreißer" in Bezug auf eine Verteilung genannt, wenn er sich in einem geeignet definierten Bereich befindet, der unter dieser Verteilung eine sehr kleine Wahrscheinlichkeit hat. Sei $\mathcal{P}_0$ eine Menge von Verteilungen auf $(M, \mathcal{B})$. Nach der Terminologie von Davies und Gather ist eine α-Ausreißerregion zu $P \in \mathcal{P}_0$ einfach eine Menge $A_\alpha \in \mathcal{B}$ mit $P(A_\alpha) = \alpha$, wobei α sehr klein sein soll. Diese Menge soll weiterhin die Vorstellung widerspiegeln, daß ihre Elemente Ausreißer sind, sie sollte also die Bereiche enthalten, in denen P am wenigsten „dicht" ist. Falls P eine unimodale Lebesgue-Dichte p hat, geben Davies und Gather (1989)

$$A_\alpha := A(\max\{c : P\{A(c) \leq \alpha\}\}) \text{ mit } A(c) := \{x : p(x) < c\}$$

an. Allgemeiner:

Definition 7.1 (Ausreißerregion) *Gegeben sei eine Abbildung*

$$A : [0,1] \times \mathcal{P}_0 \mapsto \mathcal{B}, \quad (\alpha, P) \mapsto A(\alpha, P) \text{ mit } P(A(\alpha, P)) \leq \alpha.$$

Dann heißt das Bild von $P \in \mathcal{P}_0$ unter $A(\alpha, \bullet)$ eine α-Ausreißerregion[10] *zu P.*

Die Definition der Abbildung A hängt von der anschaulichen Vorstellung ab, die man vom Begriff „Ausreißer" hat.

Um einen Ausreißer in Bezug auf einen Datensatz $\mathbf{Z} = (z_1, \ldots, z_n)' \in M^n$, $n \in \mathbb{N}$ zu ermitteln, wird nun die Ausreißerregion anhand des Datensatzes geschätzt. Die Schätzung einer Ausreißerregion zu $\mathbf{Z}$ ist also $A_n(\mathbf{Z})$, wobei $A_n : M^n \mapsto \mathcal{B}$.

Definition 7.2 (Ausreißeridentifizierer) *Sei $A_n : M^n \mapsto \mathcal{B}$ Schätzung einer Ausreißerregion. Dann ist*

$$\begin{aligned} I_n : M^n &\mapsto \{0,1\}^M, \\ \mathbf{Z} &\mapsto I_n[\mathbf{Z}] : M \mapsto \{0,1\}, \; m \mapsto 1(m \in A_n(\mathbf{Z})) \end{aligned}$$

ein Ausreißeridentifizierer, d.h. $m \in M$ wird als Ausreißer identifiziert, wenn $I_n[\mathbf{Z}](m) = 1$.

Zu einer gegebenen Indikatorfunktion I_n kann umgekehrt eine Ausreißerregionsschätzung

$$A_n : M^n \mapsto \{0,1\}^M, \; \mathbf{Z} \mapsto \{m \in M : I_n[\mathbf{Z}](m) = 1\}$$

definiert werden. Bei Davies und Gather (1989) ist ein Ausreißeridentifizierer zusätzlich abhängig von α und wird für $\alpha \in (0,1)$ definiert, d.h. ein Identifizierer dort entspricht einer Familie von Identifizierern bei mir. Für die Zwecke der Clusteranalyse wird es

[10] Davies und Gather geben im stetigen Lokationsfall die Bedingung $P(A(\alpha, P)) = \alpha$ statt „$\leq$". Ich benutze hier aber eine allgemeinere Formulierung, um auch unstetige Verteilungen mit beliebigem α behandeln zu können.

reichen, einen Identifizierer für ein bestimmtes, fest vorgegebenes α zu haben. Daher betreibe ich hier weniger Aufwand. Weiterhin werden bei Davies und Gather nur die Punkte aus $\mathbf{Z}$ und nicht alle Punkte aus M klassifiziert.

Wann kann A_n als sinnvolle Schätzung für eine Ausreißerregion betrachtet werden? Davies und Gather geben die Forderung

$$\forall P \in \mathcal{P}_0 : \; P^n\{A_n(\mathbf{Z}) \subseteq A(\alpha_n, P)\} \stackrel{!}{=} 1 - \alpha_0 \tag{7.2}$$

an[11], wobei α_0 klein sein soll. Sie betrachten also die Folge (A_n) als Folge von Schätzungen von Ausreißerregionen mit unterschiedlichem α_n. Die Wahrscheinlichkeit, daß Punkte, die nicht in $A(\alpha_n, P)$ liegen, als Ausreißer klassifiziert werden, soll dann klein sein. Für die Wahl der α_n wird

$$\begin{aligned} \forall P \in \mathcal{P}_0 : \; P^n(\{\mathbf{Z} : \{z_1, \ldots, z_n\} \subset M \setminus A(\alpha_n, P)\}) &\stackrel{!}{=} 1 - \alpha_0 \\ \Leftrightarrow \alpha_n = 1 - (1 - \alpha_0)^{\frac{1}{n}} & \end{aligned} \tag{7.3}$$

vorgeschlagen. Das bedeutet: Die Wahrscheinlichkeit, daß kein Punkt eines beobachteten Datensatzes in der Ausreißerregion ist, wenn der Datensatz eine unabhängig identisch verteilte Stichprobe aus P ist, soll mindestens $1 - \alpha_0$ (sehr groß) sein. Umgekehrt wird also, wenn ein Punkt der Stichprobe als Ausreißer identifiziert wird, der Schluß naheliegen, dieser Punkt sei nicht von P erzeugt worden. Das ist der Zusammenhang zum Ansatz mit Modellen der Form (7.1). Ein guter Identifizierer nach der Philosophie von Davies und Gather ist also nicht einer, der eine Ausreißerregion gut approximiert, wenn alle Punkte nach P erzeugt würden. Stattdessen geht es darum, in Anwesenheit von Punkten aus P und extremen anderen Punkten, die Punkte, die nach P erzeugt wurden, gut von den anderen trennen zu können.

Ich werde mit den obigen Definitionen weiterarbeiten (abgesehen von den Forderungen (7.2) und (7.3)), denn für die Idee der Fixpunktcluster ist eine Ausreißeridentifikation vonnöten, die von einem Datensatz abhängt. Sie wird dann auf den Teil des Gesamtdatensatzes angewendet, bei dem interessiert, ob er ein Cluster ist. Es sollen dann sowohl die Punkte dieses Teiles als auch die anderen Punkte klassifiziert werden können. Es wird also eine Bereichsschätzung auf M benötigt. Seien im folgenden $\mathbf{Z}, M, A, A_n, I_n$ wie gehabt bzw. in den Definitionen 7.1 und 7.2. Die restlichen Bezeichnungen sind aus Abschnitt 1.7.

Definition 7.3 (Allgemeine Fixpunktclustervektoren) *Seien $I_1, \ldots, I_n$ Ausreißeridentifizierer. Sei weiterhin*

$$f : \{0,1\}^n \mapsto \{0,1\}^n, \; g \mapsto \left(1 - I_{n(g)}[\mathbf{Z}(g)](z_1), \ldots, 1 - I_{n(g)}[\mathbf{Z}(g)](z_n)\right).$$

Dann heißt g mit $n(g) > 0$ Fixpunktclustervektor (FPCV) bzgl. $\mathbf{Z}$, wenn g Fixpunkt von f ist, also $f(g) = g$.

[11] In (7.2) und (7.3) könnte „=" durch „≥" ersetzt werden, um größere Allgemeinheit zu erreichen.

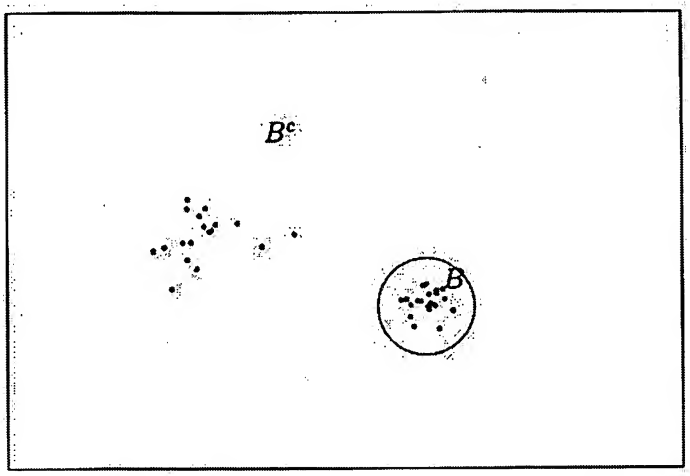


Abbildung 7: Fixpunktcluster

Zur Interpretation: Wir interessieren uns dafür, ob ein bestimmter Teildatensatz $\mathbf{Z}(g)$ von $\mathbf{Z}$ ein FPC ist. g_j ist 1, wenn z_j zum Teildatensatz $\mathbf{Z}(g)$ gehört. Definition 7.3 besagt, daß dazu für jeden Punkt z_i, $i = 1, \ldots n$ der Ausreißeridentifizierer $I_{n(g)}[\mathbf{Z}(g)](z_i)$ berechnet werden muß. Das heißt: Es muß ermittelt werden, ob z_i bezüglich $\mathbf{Z}(g)$ ein Ausreißer ist (d.h. $I_{n(g)}[\mathbf{Z}(g)](z_i) = 1$) oder nicht. g ist genau dann ein FPCV, wenn kein Punkt aus $\mathbf{Z}(g)$ Ausreißer bzgl. $\mathbf{Z}(g)$ ist, aber alle Punkte aus $\mathbf{Z}(1-g)$.

Ist g ein FPCV, so wird dadurch auch eine zugehörige geschätzte Ausreißerregion $A_{n(g)}(\mathbf{Z}(g))$ festgelegt. Die Definition der Fixpunktcluster hängt also entscheidend von der Definition der Ausreißeridentifizierer bzw. -regionen ab.

Abbildung 7 zeigt den anschaulichen Fall der Lageschätzung im $\mathbb{R}^2$, der in dieser Arbeit sonst nicht weiter behandelt wird. Sei $g(z) := (1[z_i \in B])_{i=1,\ldots,n}$. Um zu ermitteln, ob die Punkte in B ein FPC sind, muß die geschätzte Ausreißerregion $A_{n(g)}(\mathbf{Z}(g))$ ermittelt werden. In diesem Beispiel sei das B^c. Also ist kein Punkt aus B Ausreißer bzgl. $\mathbf{Z}(g)$, aber alle Punkte aus B^c. Daher ist g FPCV bzgl. $\mathbf{Z}$.

Interpretiert man die $A_{n_0}, \ldots, A_n$ als Schätzer für ein $A(\alpha, P)$ mit festem α (im Unterschied zur Konzeption von Davies und Gather), so ist der „Clusterindikator" $1 - I_{n(g)}[\mathbf{Z}(g)]$ ein natürlicher Schätzer für folgenden „Parameter" der zugrundeliegenden Verteilung:

Definition 7.4 (Allgemeine Fixpunktclusterindikatoren) *Sei $\mathcal{P}$ die Menge der Verteilungen auf $(M, \mathcal{B})$, $\mathcal{P}_0 \subseteq \mathcal{P}$. Weiterhin sei $I := \left\{g \in \{0,1\}^M : \{g = 1\} \in \mathcal{B}\right\}$ die Menge der $\mathcal{B}$-meßbaren Indikatorfunktionen. A sei Ausreißerregion mit $\alpha > 0$ fest vorgegeben,*

$$A_{\mathcal{P}} : \mathcal{P} \mapsto \mathcal{B},\ P \mapsto A_{\mathcal{P}}(P),\ \text{wobei } \forall Q \in \mathcal{P}_0 :\ A_{\mathcal{P}}(Q) = A(\alpha, Q).$$

Für $P \in \mathcal{P}$, $B \in \mathcal{B}$ sei $P_B := P(\bullet|B) \in \mathcal{P}$ (beliebig, wenn $P(B) = 0$). Nun sei

$$f : I \mapsto I,\ g \mapsto 1 - 1\left(\bullet \in A_{\mathcal{P}}(P_{\{g=1\}})\right).$$

Dann heißt g mit $P(\{g = 1\}) > 0$ Fixpunktclusterindikator (FPCI) bzgl. P, wenn $f(g) = g$.

Zur Interpretation: $\mathcal{P}_0$ enthalte die Verteilungen, die als „clustergenerierend" interpretiert werden sollen. Damit meine ich, daß Daten, die von einem gemeinsamen $Q \in \mathcal{P}_0$ erzeugt werden, in unserem Clusterverständnis „zusammengehörig" sein sollen.[12] Zum Beispiel kann $\mathcal{P}_0$ die Menge der p–dimensionalen Normalverteilungen sein, während Mischungen von Normalverteilungen oder bimodale Verteilungen eher mehrere Cluster generieren. Für jede Verteilung aus $\mathcal{P}_0$ soll eine Ausreißerregion $A(\alpha, \bullet)$ definiert sein. Diese Definition wird möglichst sinnvoll (zum Beispiel stetig oder durch direkte Übertragung der Definitionsgleichung) auf ganz $\mathcal{P}$ fortgesetzt. Die Verteilungen aus $\mathcal{P}_0$ bestimmen also den Begriff „Ausreißer" auf ganz $\mathcal{P}$.

f ordnet g die Indikatorfunktion der Menge zu, die die Punkte enthält, die bzgl. $P_{\{g=1\}}$ keine Ausreißer sind. $P_{\{g=1\}}$ ist die bedingte Verteilung P unter $\{g = 1\}$, also eine durch g abgeschnittene („truncated") Verteilung. g ist FPCI, wenn genau diese Punkte aus M nicht in der durch sie definierten Ausreißerregion liegen. In diesem Sinne sind dann die Punkte aus $\{g = 1\}$ „zusammengehörig".

Wäre $g = 1(\bullet \in B)$ in Abbildung 7, so ist g FPCI bzgl. Q, wenn $A_{\mathcal{P}}(Q_B) = B^c$, also die Punkte aus B^c genau die Ausreißer bzgl. Q_B sind. Die Punkte aus B können dann als „Fixpunktcluster" bezeichnet werden. Das kann zum Beispiel erfüllt sein, wenn Q eine zweidimensionale Dichte hat, die in der Mitte von B ein lokales Maximum hat, in der Umgebung von B sehr niedrig ist und erst weit von B entfernt wieder höhere Werte annimmt. Eine solche Verteilung hat auch den abgebildeten Datensatz erzeugt.

Ich werde nun einige Aspekte des Fixpunktclusterkonzeptes diskutieren. Definition 7.4 beinhaltet keinen direkten Zusammenhang zwischen den FPCI und den Ausreißerregionen der Verteilungen $Q \in \mathcal{P}_0$. Ist g FPCI bzgl. Q, so ist $\{g = 0\}$ nicht notwendig α–Ausreißerregion von Q nach Definition 7.1. Angenommen, es gäbe α_1 nicht notwendig gleich α, so daß ein FPCI bzgl. Verteilungen aus $\mathcal{P}_0$ Indikatorfunktion für das Komplement einer α_1-Ausreißerregion wäre, dann wäre für Q nur genau ein FPCI vorhanden, nämlich die Indikatorfunktion der Menge der α_1-Nichtausreißer. Das wäre sinnvoll, da $Q \in \mathcal{P}_0$ ja „zusammengehörige" Punktmengen generieren soll.

[12] Aufgrund der etwas mißverständlichen Begriffsbildung sei ausdrücklich betont, daß eine „clustergenerierende Verteilung" nicht etwa eine Verteilung ist, die mehrere Cluster generiert. Stattdessen soll eine „clustergenerierende Verteilung" einen homogenen Datensatz (Cluster) modellieren.

Definition 7.5 (Ausreißereigenschaft) *$\mathcal{P}_0$ hat die Ausreißereigenschaft bzgl. $A_{\mathcal{P}}$:⇔*

$$\exists \alpha_1 \in [0,1] : Q \in \mathcal{P}_0,\ g \text{ FPCI bzgl. } Q \Rightarrow A_{\mathcal{P}}(Q_{\{g=1\}}) = A(\alpha_1, Q).$$

Im hier behandelten Regressionsfall wird die Ausreißereigenschaft erfüllt sein (Bemerkung 12.3).

Der Zusammenhang zwischen Ausreißererkennung und der Analyse von Mischmodellen wie in Abschnitt 2 besteht darin, daß Mischmodelle auch die Form (7.1) haben. Wie in der auf dieser Formulierung basierten Ausreißeranalyse besteht dabei auch das Problem, daß die verschiedenen Mischungskomponenten nicht notwendig Punkte generieren, die sich mit großer Wahrscheinlichkeit deutlich voneinander unterscheiden (vgl. zum Beispiel Abbildung 2 in der Einleitung). Bei der stochastischen Clusteranalyse durch Mischmodelle ist also zu beachten, daß ein Mischmodell nur dann verschiedene Cluster im anschaulichen Sinne generiert, wenn die Mischungskomponenten hinreichend unterschiedlich sind. Der Idee nach sollen FPC Cluster in einem anschaulichen Sinne sein. Das Ziel ist also nicht in erster Linie, die Parameter eines Mischmodells zu schätzen, sondern Klumpen von Daten zu finden.

In dieser Arbeit wird es fast ausschließlich um die Anwendung des FPC-Konzeptes auf den Fall linearer Regression gehen. Nur im folgenden Abschnitt werde ich kurz die Anwendung der Idee auf eine andere Situation skizzieren.

Festzuhalten ist, daß auch allgemeine Fixpunktcluster die bereits in Abschnitt 3.5.1 erklärten Eigenschaften „lokale Definition eines Clusters", „kein Partitionszwang" und „keine Optimallösung eines Entscheidungsproblems" haben.

Ein Fixpunktcluster wird immer durch einen Indikator g definiert. Das Konzept ist auch in dem Sinne „lokal", daß Fixpunktcluster eines Datensatzes oder einer Verteilung erhalten bleiben, wenn Punkte (bzw. eine weitere Verteilung) hinzugefügt oder entfernt werden, die außerhalb des Clusterbereiches $\{g = 1\}$ liegen, also Ausreißer bzgl. des Clusters sind:

Korollar 7.6 (Fixpunktcluster und Ausreißer) *Sei $\mathbf{Z}_1 := (z_1, \ldots, z_{n_1})'$, $\mathbf{Z}_2 := (z_1, \ldots, z_{n_2})'$ mit $n_2 > n_1 > 0$. Seien $I_1, \ldots, I_{n_2}$ Ausreißeridentifizierer. Sei für $g \in \{0,1\}^{n_1}$*

$$\forall n_2 \geq i > n_1 :\ I_{n(g)}[\mathbf{Z}_1(g)](z_i) = 1.$$

Dann ist g FPCV bzgl. $\mathbf{Z}_1$ genau dann, wenn $h = (g_1, \ldots, g_{n_1}, 0, \ldots, 0)$ FPCV bzgl. $\mathbf{Z}_2$ ist.

Seien $P, R \in \mathcal{P}$ und für $g_P \in I$ sei $R_{\{g_P=1\}} = P_{\{g_P=1\}}$. Dann ist g_P FPCI bzgl. R und $P(\{g_P = 1\}) > 0$ genau dann, wenn g_P FPCI bzgl. P ist und $R(\{g_P = 1\}) > 0$.

Beweis: Bezeichne $f_{\mathbf{Z}_1}$ und $f_{\mathbf{Z}_2}$ die Funktionen f aus Definition 7.3 bzgl. $\mathbf{Z}_1$ und $\mathbf{Z}_2$.

Offenbar ist $\mathbf{Z}_1(g) = \mathbf{Z}_2(g)$. Es folgt direkt

$$g = f_{\mathbf{Z}_1}(g) \Leftrightarrow h = f_{\mathbf{Z}_2}(h).$$

Analog bezeiche f_P und f_R die Funktionen f aus Definition 7.4 bzgl. P und R. Dann folgt direkt

$$g_P = f_P(g_P) \Leftrightarrow g_P = f_R(g_P).$$

In Abbildung 7 klumpen sich die Punkte im Zentrum von B. Das Korollar besagt hier, daß diese Punkte ein FPC bleiben, wenn in B^c Punkte hinzugefügt oder weggenommen werden.

Die „lokale" Definition des FPC bringt die Möglichkeit, von einzelnen vorgegebenen Teildatensätzen von Interesse direkt auszurechnen, ob sie FPC sind. Andererseits ist es schwierig, die Fixpunktcluster eines gegebenen Datensatzes bzw. einer gegebenen Verteilung vollständig zu bestimmen. Theoretisch müßte man die Fixpunktclustereigenschaft für jeden Indikatorvektor aus $\{0,1\}^n$ bzw. jede meßbare Indikatorfunktion auf M einzeln nachprüfen, was normalerweise kaum möglich sein wird. In Teil III werden zum Beispiel im Regressionsfall zu verschiedenen Verteilungen FPCI berechnet. Bis auf eine Ausnahme werden dort nur Existenzen, aber keine Eindeutigkeiten gezeigt.

Unter Umständen können aber relevante und irrelevante Fixpunktcluster unterschieden werden: Im Falle von Daten aus einer Mischverteilung können zum Beispiel die FPC „relevant" genannt werden, die den Mischungskomponenten entsprechen, wobei noch zu präzisieren wäre, was das bedeutet. Die Problemstellung wäre dann, unter Zuhilfenahme eines geeigneten Algorithmus (für den Regressionsfall siehe Abschnitt 9) alle relevanten FPCV zu finden. Inwiefern das gelingt, wird in den Simulationen in Teil IV untersucht.

Im Falle eines Datensatzes aus einer Verteilung mit nicht identifizierbaren Parametern - aber deutlich getrennten Mischungskomponenten - kann ein Verfahren mit einer „lokalen" Clusterdefinition in der Lage sein, alle unterschiedlichen Möglichkeiten zu finden. In Abschnitt 16.2 wird unter anderem eine solche Situation simuliert.

7.2 Beispiel: Fixpunktcluster für 0-1-Vektoren

Bevor ich zum linearen Regressionsfall komme, gebe ich ein alternatives Beispiel für die Anwendung des Fixpunktcluster-Konzeptes. Damit soll zum einen die Idee illustriert werden. Zum anderen soll gezeigt werden, daß der FPC-Ansatz auch über den Regressionsfall hinaus mit Gewinn angewendet werden kann. Sei in diesem Abschnitt $\mathbf{Z} = (z_1, \ldots, z_n)'$, $z_i = (z_{i1}, \ldots, z_{ik})' \in \{0,1\}^k$ für $i = 1, \ldots, n$. Es geht also darum, Cluster aus einem Datensatz von k-dimensionalen 0-1-Vektoren zu bilden. Es sei $\bar{z}_j := \frac{1}{n}\sum_{i=1}^n z_{ij}$ für $j = 1, \ldots, k$. $B[k,p]$, $k \in I\!N$, $p \in [0,1]$ bezeichne die Binomial(k,p)-Verteilung (bzw. deren VF), $AB[k,p_1,\ldots,p_k]$, $k \in I\!N$, $p_1, \ldots, p_k \in [0,1]$ bezeichne die verallgemeinerte Binomialverteilung mit Parametern $(k, p_1, \ldots, p_k)$,[13]

$$P[k,p_1,\ldots,p_k] := \bigotimes_{j=1}^{k} B[1,p_j], \quad \mathcal{P}_0 := \left\{P[k,p_1,\ldots,p_k] : (p_1,\ldots,p_k) \in [0,1]^k\right\}$$

[13]Unter der „verallgemeinerten Binomialverteilung" mit Parametern $k, p_1, \ldots, p_k$ verstehe ich die Verteilung einer Summe aus k unabhängig $B[1,p_i]$-verteilten Zufallsvariablen, $i = 1, \ldots, k$.

für gegebenes k. Die Menge $\mathcal{P}_0$ sei die Menge der „clustergenerierenden" Verteilungen wie in Definition 7.4. Das bedeutet: Wir betrachten 0-1-Vektoren als „zusammengehörig", wenn sie aussehen, als seien sie von derselben Kopplung unabhängiger Bernoulliverteilungen erzeugt. Anschaulich könnte man nun einen Punkt als Ausreißer zu einer gegebenen Verteilung $P[k, p_1, \ldots, p_k] \in \mathcal{P}_0$ bezeichnen, wenn möglichst viele seiner k Komponenten von der zu erwartenden Mehrheit abweichen. Formal:

$$\forall P[k, p_1, \ldots, p_k] \in \mathcal{P}_0 :$$
$$A(\alpha, P[k, p_1, \ldots, p_k]) := \left\{z \in \{0,1\}^k : m[p_1, \ldots, p_k](z) \geq c(\alpha)\right\},$$
$$\text{wobei } c(\alpha) := \min\left\{c : 1 - B[k, \tfrac{1}{2}](c) \leq \alpha\right\},$$
$$m[p_1, \ldots, p_k](z) := \sum_{j=1}^{k} \left[z_j 1\left(p_j \leq \frac{1}{2}\right) + (1 - z_j) 1\left(p_j > \frac{1}{2}\right)\right].$$

$m[p_1, \ldots, p_k](z)$ ist also ein Zähler dafür, wie häufig z_j zur „Minderheit" gehört. Im Falle $p_j = \frac{1}{2}$ ist das etwas willkürlich, es könnte auch „$<$" und „$\geq$" heißen. Indem $m[p_1, \ldots, p_k]$ durch $m[\int z_1 dP(z), \ldots, \int z_k dP(z)]$ ersetzt wird, kann die Definition von A kanonisch auf alle Verteilungen P auf $\{0,1\}^k$ ausgedehnt werden.

Bemerkung 7.7 *Die Abbildung A definiert für $P \in \mathcal{P}_0$ tatsächlich eine α-Ausreißerregion nach Definition 7.1. Es gilt nämlich:*

$$\mathcal{L}(z) = P[k, p_1, \ldots, p_k] \Rightarrow \mathcal{L}(m[p_1, \ldots, p_k](z)) = AB[k, q_1, \ldots, q_k],$$
$$\text{wobei für } j = 1, \ldots, k\text{: } q_j = p_j 1(p_j \leq \tfrac{1}{2}) + (1 - p_j) 1(p_j > \tfrac{1}{2}) \leq \tfrac{1}{2}.$$

Weiter gilt:

$$AB[k, q_1, \ldots, q_k]\{y : y \geq c(\alpha)\} \leq B\left[k, \frac{1}{2}\right]\{y : y \geq c(\alpha)\} \leq \alpha. \tag{7.4}$$

Gleichung (7.4) folgt aus Theorem A auf S.31 von Szekli (1995). Dieser Satz besagt, daß die Verteilung einer Summe von unabhängigen nichtnegativen[14] Zufallsvariablen sich stochastisch vergrößert, wenn einzelne Summanden durch Zufallsvariablen mit stochastisch größerer Verteilung ersetzt werden.

Für $k = 5$, $\alpha = 0.05$ gilt zum Beispiel $c(\alpha) = 5$, denn $1 - B[5, \frac{1}{2}](5) = \frac{1}{32}$ und $1 - B[5, \frac{1}{2}](4) > 0.05$. Mit $P = P[k, 0, 0, \frac{1}{3}, \frac{2}{3}, \frac{2}{3}]$ ist zum Beispiel $A(\alpha, P) = \{(1,1,1,0,0)\}$.

Bemerkung 7.8 *$g := 1(\bullet \in \{0,1\}^5 \setminus \{(1,1,1,0,0)\})$ ist in diesem Beispiel FPCI bzgl. P, denn $P\{(1,1,1,0,0)\} = 0$, also $P(\bullet|\{(1,1,1,0,0)\}^c) = P$, $A_P(P_{\{(1,1,1,0,0)\}^c}) = A(\alpha, P)$ und $f(g) = 1(\bullet \in A(\alpha, P)^c) = g$ nach Definition 7.4. Für dieses spezielle P und g gilt damit auch die rechte Seite der Gleichung aus Definition 7.5 mit $\alpha_1 = \alpha$. Für die Erfüllung der Ausreißereigenschaft müßte für $P \in \mathcal{P}_0$ zusätzlich gezeigt werden, daß es keine weiteren FPCI gibt.*

[14] Diese Einschränkung wäre nicht nötig und wird von Szekli gemacht, weil der Satz zusätzlich eine Aussage über die Zahl der Summanden enthält.

Der kanonische Schätzer der Ausreißerregion $A(\alpha, \bullet)$ und der zugehörige Ausreißeridentifizierer sind dann

$$A_n(\mathbf{Z}) := \{z : m(\bar{z}_1, \ldots, \bar{z}_k) \geq c(\alpha)\},\ I_n[\mathbf{Z}](z) := 1(z \in A_n(\mathbf{Z})).$$

Für $k = 5, \alpha = 0.05$ sei zum Beispiel $n = 4$, $z_1 = (0,0,0,1,1), z_2 = (0,0,0,0,1), z_3 = (0,0,1,1,0), z_4 = (1,1,1,0,0)$. Dann ist $g = (1,1,1,0)$ FPCV bzgl. $\mathbf{Z}$. Es ist nämlich

$$\mathbf{Z}(g) = (z_1, z_2, z_3),$$
$$\overline{z(g)}_1 = 0,\ \overline{z(g)}_2 = 0,\ \overline{z(g)}_3 = \tfrac{1}{3},\ \overline{z(g)}_4 = \tfrac{2}{3},\ \overline{z(g)}_5 = \tfrac{2}{3},$$
$$m := m[\overline{z(g)}_1, \ldots \overline{z(g)}_k],\ m(z_1) = 0,\ m(z_2) = 1,\ m(z_3) = 1,\ m(z_4) = 5,$$
$$1 - I_{n(g)}[\mathbf{Z}(g)](z_i) = 1 \text{ für } i = 1,2,3,\ 1 - I_{n(g)}[\mathbf{Z}(g)](z_4) = 0,$$

d.h. bzgl. $\mathbf{Z}(g)$ ist nur z_4 Ausreißer und $f(g) = g$ gemäß Definition 7.3. Ist dagegen $g = (1,0,1,1)$, so berechnet man

$$\overline{z(g)}_1 = \tfrac{1}{3},\ \overline{z(g)}_2 = \tfrac{1}{3},\ \overline{z(g)}_3 = \tfrac{2}{3},\ \overline{z(g)}_4 = \tfrac{2}{3},\ \overline{z(g)}_5 = \tfrac{1}{3},$$

damit $m(z_2) = 3 < 5$ und daher $1 - I_{n(g)}[\mathbf{Z}(g)](z_2) = 1 \neq g_2$. z_1, z_3 und z_4 bilden also keinen Cluster, da z_2 kein Ausreißer ist.

7.3 Fixpunktcluster und die Selbstorganisation der Wahrnehmung

Ich werde kurz einen nichtstochastischen Zugang zur Fixpunktcluster-Idee erläutern. Fixpunktcluster können als Modellierung von Objekten der Kognition im Rahmen der Selbstorganisationstheorie interpretiert werden.

Heinz von Förster (1976) formalisiert menschliche kognitive Aktionen als rekursive Operatoren auf einem Bereich von „Observablen". Das heißt: Eine kognitive Handlung operiert mit ihrem eigenen Ergebnis:

$$\text{obs}_{i+1} = \text{COORD}(\text{obs}_i). \tag{7.5}$$

„COORD" steht für „Koordination". Kognitive Prozesse werden nun als die rekursive Berechnung eine Realität abhängig von der Weise, in der das Subjekt operiert, aufgefaßt. Von Förster (1973) [15] gibt dazu folgende Illustration:

Kognition ⟶ *Berechnungen von* ⟲

Dieser Ansatz kann auch mit biologischen Argumenten begründet werden. Von Förster verweist dafür auf Piaget (1975) und Maturana (1970).

Von Förster erklärt, daß für eine externe Beobachter/in die „Objekte" einer Person nicht von den Fixpunkten[16] der Operation (7.5) unterschieden werden können. Das bedeutet: Die Beobachter/in nennt ein Objekt „erkannt" durch die beobachtete Person, wenn die Koordination der Handlungen der Person bezüglich des Objektes stabil sind.

[15] Die deutsche Übersetzung beider Arbeiten ist enthalten in von Förster (1993).

[16] Von Förster benutzt dafür den älteren Ausdruck „Eigenwert".

Ein solcher Prozeß wird durch Fixpunktclusterbildung modelliert. In dieser Terminologie wären die Cluster (oder Muster) eines Datensatzes die zu erkennenden Objekte. Stellen wir uns einen Fixpunktalgorithmus $g^{i+1} = f(g^i)$ vor. g^0 definiert dann den Teildatensatz, mit dem die Iteration beginnt, die initiale Observable. Die Funktion f definiert die Handlungen, die der Algorithmus mit den Daten durchführt, und ein FPC ist ein stabiler Punkt dieses Prozesses.

8 Fixpunktcluster im Regressionsfall

8.1 Regressions-Fixpunktclusterindikatoren

Das für meine Zwecke am einfachsten handhabbare Modell aus Abschnitt 2 ist das Mischmodell 3 mit stochastischen Regressoren und Normalverteilungsannahme für den Störterm. In der späteren Theorie werde ich mich auf dieses Modell beschränken. Deshalb wird auch bei der Einführung der FPCI für lineare Regression nur der Fall stochastischer Regressoren betrachtet. Man käme aber mit festen Regressoren zur selben Definition des FPCV für Datensätze (gemäß Abschnitt 8.2). Sei also mit den Bezeichnungen aus Modell 3

$$\mathcal{P}_0 := \{F(\bullet, \beta, \sigma^2, G) : (\beta, \sigma^2, G) \in T_s, E_G(xx') \text{ invertierbar}, E_G(\|x\|^2) < \infty\} \quad (8.1)$$

die Menge der clustergenerierenden Verteilungen. Ein Punkt kann sinnvollerweise ein Ausreißer bzgl. einer Regressionsverteilung $F(\bullet, \beta, \sigma^2, G)$ genannt werden, wenn er weit von der durch β definierten Regressionshyperebene entfernt ist. Also:

$$A(\alpha, F(\bullet, \beta, \sigma^2, G)) := \left\{(x, y) : (y - x'\beta)^2 > c(\alpha)\sigma^2\right\} \forall F(\bullet, \beta, \sigma^2, G) \in \mathcal{P}_0. \quad (8.2)$$

wobei $c(\alpha)$ das $(1-\alpha)$–Quantil der χ^2_1-Verteilung sei, so daß $P(A(\alpha, P)) = \alpha \, \forall P \in \mathcal{P}_0$.

Bemerkung 8.1 *Man könnte auch Ausreißerregionen definieren, die von der Verteilung der Regressoren abhängig sind. Im Rahmen dieser Arbeit sollen aber Daten mit deutlich unterschiedlichen Regressoren, aber gleichem linearen Zusammenhang als zusammengehörig betrachtet werden.*

A kann folgendermaßen auf $\mathcal{P} := \{P \in \mathcal{P}_{p+2} : \mathcal{L}(X_{p+1}) = \delta_1\}$ erweitert werden:

$$\forall P \in \mathcal{P}: \; A_{\mathcal{P}}(P) := \{(x, y) : (y - x'\beta(P))^2 > c(\alpha)\sigma^2(P)\}, \text{ wobei}$$
$$\beta(P) := \arg\min_{\beta} \int (y - x'\beta)^2 dP(x, y),$$
$$\sigma^2(P) := \int (y - x'\beta(P))^2 dP(x, y),$$

falls $\arg\min_{\beta} \int (y - x'\beta)^2 dP(x, y)$ existiert und eindeutig ist, und $A_{\mathcal{P}}(P) = \emptyset$ sonst. $\beta(P)$ ist die Verallgemeinerung des KQ-Schätzers auf W-Maße. Ein eindeutiges $\beta(P)$ existiert zum Beispiel, wenn $E(\|x\|^2)$, $E(y^2)$ und $[E(xx')]^{-1}$ existieren; siehe Hilfssatz 11.1.

Bemerkung 8.2 *Diese Erweiterung ist nicht die einzig mögliche und vermutlich nicht einmal eine besonders gute; da sie durch die Verwendung des KQ-Funktionals nicht schwach stetig und damit nicht qualitativ robust ist (siehe zum Beispiel Huber (1981), S. 10). Ich werde in dieser Arbeit aber dabei bleiben, da sie am einfachsten handhabbar ist.*

Damit ergibt sich aus Definition 7.4:

Definition 8.3 (KQ-Fixpunktcluster-Indikatoren) *Sei $P \in \mathcal{P}$ und $g : \mathbb{R}^{p+1} \mapsto \{0,1\}$ eine $\mathcal{B}^{p+1}$-meßbare Indikatorfunktion*[17]*. Für gegebenes $c > 0$ („Fixpunktcluster-Justierkonstante") ist g ein KQ-Fixpunktcluster-Indikator (KQ-FPCI) bzgl. $P :\Leftrightarrow$*

$$\int g(x,y)dP(x,y) > 0, \tag{8.3}$$

$$\arg\min_{\beta} \int (y - x'\beta)^2 g(x,y)dP(x,y) \text{ existiert und ist eindeutig,} \tag{8.4}$$

$$g(x,y) = 1\left[(y - x'\beta(g,P))^2 \le c\sigma^2(g,P)\right] \ \forall (x,y) \in \mathbb{R}^p \times \{1\} \times \mathbb{R}, \tag{8.5}$$

$$\text{wobei } \beta(g,P) := \arg\min_{\beta} \int (y - x'\beta)^2 g(x,y)dP(x,y), \tag{8.6}$$

$$\sigma^2(g,P) := \frac{\int (y - x'\beta(g,P))^2 g(x,y)dP(x,y)}{\int g(x,y)dP(x,y)}. \tag{8.7}$$

In Definition 8.3 werden KQ-FPCI über die Fixpunktgleichung (8.5) für Indikatorfunktionen definiert. Für die spätere Theorie ist es manchmal nützlicher, eine äquivalente Formulierung mit einer Fixpunktgleichung für die Regressionsparameter zu benutzen:

Bemerkung 8.4 *Sei $c > 0$ fest, $Q \in \mathcal{P}$. Für $\theta \in \mathbb{R}^{p+1}$, $s^2 \in \mathbb{R}_0^+$ sei $g_{\theta,s^2} : \mathbb{R}^{p+1} \mapsto \{0,1\}$ definiert gemäß*

$$g_{\theta,s^2}(x,y) := 1\left((y - x'\theta)^2 \le cs^2\right). \tag{8.8}$$

Dann ist g_{θ,s^2} nach Definition 8.3 KQ-FPCI bzgl. Q genau dann, wenn (8.3) und (8.4) für $g = g_{\theta,s^2}$ erfüllt sind und $(\theta, s^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ Fixpunkt ist von $f : \mathbb{R}^{p+1} \times \mathbb{R}_0^+ \mapsto \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ gemäß

$$f(\theta, s^2) = \left(\beta(g_{\theta,s^2}, Q), \sigma^2(g_{\theta,s^2}, Q)\right), \tag{8.9}$$

β definiert in (8.6), σ^2 in (8.7).

Bemerkung 8.5 *Für $P \in \mathcal{P}$ sei mit D gemäß (2.2)*

$$\forall B \in \mathcal{B}^{p+2} : \ P^D(B) := P\{(x,y) : \ D(x,y) \in B\}.$$

Dann gilt: Ist g KQ-FPCI bzgl. P, so ist g^* gemäß

$$g^*(x,y) := g\left(D^{-1}(x,y)\right) \ \forall (x,y) \in \mathbb{R}^{p+2}$$

KQ-FPCI bzgl. P^D und

$$\beta(g^*, P^D) = (\Gamma^{-1})'(a\beta(g,P) + b), \quad \sigma^2(g^*, P^D) = a^2\sigma^2(g,P).$$

Also sind KQ-FPCI äquivariant gegenüber Transformationen der Form (2.2).

[17] Der Einfachheit halber schreibe ich auch im Regressionsfall mit Achsenabschnitt immer $g(x,y)$ statt $g(x^-,y)$.

Beweis: Mit der Transformationsformel gilt $\int g^* dP^D = \int g\, dP$ und damit auch (8.3) für g^* bzgl. P^D. Da D eine invertierbare affin-lineare Transformation ist, folgt auch (8.4) für g^* bzgl P^D. Weiter gilt

$$\beta(g^*, P^D) = \arg\min_\beta \int (y - x'\beta)^2 g^*(x,y) dP^D(x,y) =$$
$$= \arg\min_\beta \int (ay + x'b - x'\Gamma'\beta)^2 g(x,y) dP(x,y) =$$
$$= \arg\min_\beta \int \left(y - x'\frac{\Gamma'\beta - b}{a}\right)^2 g(x,y) dP(x,y) = (\Gamma^{-1})'(a\beta(g,P) + b) \text{ und}$$
$$\sigma^2(g^*, P^D) = \frac{\int (y - x'\beta(g^*,P^D))^2 g^*(x,y) dP^D(x,y)}{\int g^*(x,y) dP^D(x,y)} =$$
$$= \frac{\int (ay + x'b - x'\Gamma'\beta(g^*,P^D))^2 g(x,y) dP(x,y)}{\int g(x,y) dP(x,y)} = a^2 \frac{\int (y - x'\beta(g,P))^2 g(x,y) dP(x,y)}{\int g(x,y) dP(x,y)} = a^2\sigma^2(g,P).$$

Mit (8.5):

$$g^*(x,y) = g\left(\Gamma^{-1}x, \tfrac{y - (\Gamma^{-1}x)'b}{a}\right) = 1\left(\left[\tfrac{y - (\Gamma^{-1}x)'b}{a} - (\Gamma^{-1}x)'\beta(g,P)\right]^2 \le c\sigma^2(g,P)\right) =$$
$$= 1\left([y - x'(\Gamma^{-1})'(a\beta(g,P) + b)]^2 \le ca^2\sigma^2(g,P)\right) =$$
$$= 1\left([y - x'\beta(g^*,P^D)]^2 \le c\sigma^2(g^*,P^D)\right).$$

Also gilt (8.5) auch für g^* bzgl. P^D, so daß alles gezeigt ist.

8.2 Regressions-Fixpunktclustervektoren

Es ist nun naheliegend, die in (8.2) definierte Ausreißerregion zu schätzen, indem β und σ^2 einfach durch die üblichen Schätzer ersetzt werden, d.h.

$$A_n(\mathbf{Z}) := \{(x,y) : (y - x'\beta(\mathbf{Z}))^2 > c(\alpha)\sigma^2(\mathbf{Z})\}, \text{ wobei}$$
$$\beta(\mathbf{Z}) := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y,$$
$$\sigma^2(\mathbf{Z}) := \tfrac{1}{n-p-1}\textstyle\sum_{i=1}^n (y_i - x_i'\beta(\mathbf{Z}))^2; \qquad (8.10)$$

falls $\mathbf{X}'\mathbf{X}$ invertierbar und $n > p+1$, $A_n(\mathbf{Z}) := \emptyset$ sonst mit den Bezeichnungen aus Abschnitt 1.7. Eingesetzt in Definition 7.3 ergibt sich also

Definition 8.6 (KQ-Fixpunktclustervektoren) *Für festes $c > 0$ ist $g = (g_1, \ldots, g_n) \in \{0,1\}^n$ KQ-Fixpunktclustervektor (KQ-FPCV) bzgl. $\mathbf{Z}$:⇔*

$$(\mathbf{X}(g)'\mathbf{X}(g))^{-1} \text{ existiert, } n(g) > p+1, \qquad (8.11)$$
$$g = f(g), \text{ wobei } f_i(g) = 1\left([y_i - x_i'\beta(\mathbf{Z}(g))]^2 \le c\sigma^2(\mathbf{Z}(g))\right) \quad \forall i = 1, \ldots, n.$$

Falls $\mathbf{X}'\mathbf{X}$ nicht invertierbar oder $n \le p+1$, sei nur $(1, \ldots, 1)$ KQ-FPCV bzgl. $\mathbf{Z}$.

Bemerkung 8.7 *KQ-FPCV sind linear äquivariant: Zu gegebenem* $\mathbf{Z}$ *sei* $\mathbf{Z}^D$ *definiert wie in Bemerkung 3.3. Ist* g *KQ-FPCV bzgl.* $\mathbf{Z}$*, so ist* g *auch KQ-FPCV bzgl.* $\mathbf{Z}^D$*, wobei*

$$\beta(\mathbf{Z}^D(g)) = (\Gamma^{-1})'(a\beta(\mathbf{Z}(g)) + b), \quad \sigma^2(\mathbf{Z}^D(g)) = a^2\sigma^2(\mathbf{Z}(g)).$$

Beweis: *Sei* P *die empirische Verteilung zu* $\mathbf{Z}$*,* P^D *die empirische Verteilung zu* $\mathbf{Z}^D$*,*

$$g_0(x,y) := 1\left([y - x'\beta(\mathbf{Z}(g))]^2 \leq c\sigma^2(\mathbf{Z}(g))\right),$$

also $g_i = g_0(x_i, y_i)$ *für* $i = 1, \ldots, n$*.* g_0^* *sei definiert analog zu* g^* *in Bemerkung 8.5. Dann gilt auch* $g_i = g_0^*[D((x_i, y_i))]$*. Nun folgt alles aus dem Beweis von Bemerkung 8.5.*[18]

Nach den Kriterien von Davies und Gather (1993) ist eine Ausreißeridentifikation gut, wenn sie in der Lage ist, auch in Anwesenheit mehrerer ungünstig plazierter Daten noch deutliche Ausreißer zu identifizieren und andererseits nicht zu viele „gute Daten" für Ausreißer zu halten. Das bedeutet ungefähr, daß ein guter Ausreißeridentifizierer bei Daten, die gemäß (7.1) mit ungünstigem H^* erzeugt sind, die α_n-Ausreißerregion von H_0 noch gut schätzen soll. In diesem Sinne ist die Schätzung nach (8.10) schlecht, denn $\beta(\mathbf{Z})$ und $\sigma^2(\mathbf{Z})$ haben beide Bruchpunkt (breakdown point) $\frac{1}{n}$, d.h. schon ein einziger extremer Datenpunkt kann sie beliebig weit von den zu schätzenden Parametern entfernen (siehe zum Beispiel Rousseeuw und Leroy (1988)). Für die Fixpunktclusteridee kann die Schätzung aber trotzdem tauglich sein, denn falls die Fixpunktgleichung in Definition 7.3 erfüllt ist, wird die Ausreißeridentifikation nur mit einem Teildatensatz berechnet. Wenn dieser Teildatensatz nur Daten enthält, die zusammengehörig sind, d.h. keinen Ausreißer enthalten, kann die zugehörige Ausreißerregion mit (8.10) vernünftig geschätzt werden. Die Entscheidung, ob ein solcher Teildatensatz ein FPC ist, wird also aufgrund einer brauchbaren Parameterschätzung vorgenommen.

Andererseits besteht natürlich die Gefahr, daß Teildatensätze mit extremen Ausreißern ebenfalls FPC sind, weil die Ausreißer wegen der unrobusten Parameterschätzer nicht entdeckt werden. Sowohl in der Theorie als auch bei den Simulationen wird im folgenden entsprechend meistens die Existenz sinnvoller Cluster gezeigt, nur selten aber die Nichtexistenz von unsinnigen. Ich habe mich in dieser Arbeit auf die Schätzung gemäß (8.10) beschränkt, da sie mathematisch am einfachsten handhabbar ist. Die Verwendung einer robusteren Ausreißeridentifikation könnte die Resultate des Verfahrens verbessern. Als einfachste Möglichkeit bietet sich hier an, $\beta(\mathbf{Z})$ und $\sigma^2(\mathbf{Z})$ durch robuste Schätzer zu ersetzen sowie das KQ-Funktional bei den FPCI durch die entsprechenden robusteren Funktionale, siehe Bemerkung 8.2. Allerdings hätte das wesentlich höhere Rechenzeiten zur Folge, wollte man bei der Berechnung gemäß Abschnitt 9 vorgehen.

Zur Wahl der FPC-Justierkonstante c: Gemäß (8.2) sollte c ein hohes Quantil der χ^2_1-Verteilung sein. Für eine Verteilung Q aus $\mathcal{P}_0$ sollte ein FPCI g existieren, dessen Parameter $\beta(g, Q_{\{g=1\}})$ und $\sigma(g, Q_{\{g=1\}})$ nicht sonderlich von β und σ^2 abweichen. Entsprechend sollten bei einer Stichprobe aus Q möglichst wenig Punkte als Ausreißer klassifiziert werden, da Q eine clustergenerierende Verteilung sein soll und die Punkte aus Q in diesem Sinne alle zusammengehören. In Bemerkung 12.3 wird gezeigt, daß das Komplement eines FPC eine α_1-Ausreißerregion von Q mit $\alpha_1 > \alpha$ ist. Diese Überlegungen deuten darauf hin, daß α kleiner als 0.05 oder 0.01 gewählt werden sollte, da sonst mit zu

[18] Es gilt $\sigma^2(g_0^*, P^D)\frac{n(g)}{n(g)-p-1} = \sigma^2(\mathbf{Z}^D)$. Das macht im Beweis aber keinen Unterschied.

großer Wahrscheinlichkeit „gute" Punkte als Ausreißer klassifiziert würden und damit Information für die Parameterschätzung verlorenginge. Weiterhin gibt es nach meiner Erfahrung mit $\alpha \geq 0.01$ in vielen Datensätzen extrem viele FPCV. Es läge also nahe, α wie α_n in (7.3) zu wählen. Allerdings müssen mithilfe von (8.10) bei der Bestimmung von FPC zu Teildatensätzen mit unterschiedlichen n Ausreißerregionen mit gleichen α geschätzt werden.

Ich habe in den Simulationen und Beispielen mit $c = 10$ gearbeitet, was $\alpha = 1 - \chi_1^2(10) = 0.00157$ entspricht. Das entspricht α_n gemäß (7.3) für $\alpha_0 = 0.05$ und $n = 32.65$, d.h. die Wahrscheinlichkeit, daß von 32 Punkten aus Q keiner als Ausreißer klassifiziert wird, ist noch größer als 0.95. Das ist gleichzeitig die Wahrscheinlichkeit, daß ein kompletter homogener Datensatz dieser Größe aus Q ein FPC ist.

Alle theoretischen Resultate setzen untere Schranken für c voraus. Die höchste untere Schranke ist 7.4065 und wird in Satz 13.11 benötigt. Die Beispiele 13.6 und 13.14 zeigen, daß Mischungskomponenten in Mischverteilungen weniger gut getrennt sein müssen, um KQ-FPCI zu erzeugen, wenn c so klein wie möglich gewählt wird. Das stimmt mit meiner Beobachtung überein, daß es mit niedrigem c mehr FPC gibt - sinnvolle wie sinnlose.

9 Berechnung von KQ-Fixpunktclustervektoren

Ich habe bereits erwähnt, daß man theoretisch alle Teildatensätze eines gegebenen Datensatzes $\mathbf{Z}$ untersuchen müßte, um alle KQ-FPCV zu finden. Das ist natürlich unmöglich, außer bei ganz kleinem n. Um gezielt FPCV zu finden, liegt es nahe, den üblichen Fixpunktalgorithmus zu verwenden: Man startet mit $g^0 \in \{0,1\}^n$ und iteriert $g^{k+1} = f(g^k)$ bis $g^k = f(g^k)$ mit f gemäß Definition 8.6. Diese Prozedur wiederholt man sehr häufig (siehe dazu Abschnitt 15.1.1), zum Beispiel mit zufällig gewählten g^0 oder auch mit Teildatensätzen, von denen man vorher den Verdacht hat, sie seien zusammengehörig. Natürlich kann man auch dann nicht sicher sein, alle im Datensatz enthaltenen FPCV zu finden. Das Ergebnis des Verfahrens bleibt in diesem Sinne zufällig. Entscheidend ist aber letztlich, ob die am Ende gefundenen FPC sinnvolle neue Erkenntnisse über den Datensatz bringen, worüber man anhand der Simulationsergebnisse in Abschnitt 16 befinden mag[19].

Um Konvergenz des Fixpunktalgorithmus und damit auch die Existenz von KQ-FPCV für gegebene Datensätze zu zeigen, muß er etwas modifiziert werden:

Algorithmus 1: Wähle g^0 mit $n(g^0) > p+1$

Schritt 1: Berechne $\beta(\mathbf{Z}(g^k)), \sigma^2(\mathbf{Z}(g^k))$.

Schritt 2:

$$g_i^{k+1} = g_i^k + 1\left[1\left((y_i - x_i'\beta[\mathbf{Z}(g^k)])^2 \leq c\sigma^2[\mathbf{Z}(g^k)]\right) > g_i^k\right],\ i = 1,\ldots,n.$$

(g^{k+1} indiziert alle durch g^k indizierten Punkte und alle übrigen Punkte, die bzgl. g^k keine Ausreißer sind.)

Schritt 3: Berechne $\beta(\mathbf{Z}(g^{k+1})), \sigma^2(\mathbf{Z}(g^{k+1}))$.

[19] Man vergleiche die Diskussion über relevante FPC auf Seite 60.

Schritt 4:

$$g_i^{k+2} = g_i^{k+1} - 1\left[1\left((y_i - x_i'\beta[\mathbf{Z}(g^{k+1})])^2 \leq c\sigma^2[\mathbf{Z}(g^{k+1})]\right) < g_i^{k+1}\right],\ i = 1,\ldots,n.$$

(g^{k+2} indiziert nur die durch g^{k+1} indizierten Punkte, die bzgl. g^{k+1} keine Ausreißer sind.)

Schritt 5: Ende, wenn $g^k = g^{k+1} = g^{k+2}$, sonst $k = k+2$, Schritt 1.

Algorithmus 2: Ersetzt man Schritt 2 durch

$$g_i^{k+1} = 1((y_i - x_i'\beta(\mathbf{Z}(g^k)))^2 \leq c\sigma^2(\mathbf{Z}(g^k))),\ i = 1,\ldots,n, \tag{9.1}$$

und läßt Schritt 3 und 4 weg, so erhält man den üblichen Fixpunktalgorithmus, der erfahrungsgemäß auch immer konvergiert. Ein Beweis dafür wäre aber vermutlich extrem umständlich. Da dieser Algorithmus etwas schneller ist als Algorithmus 1 und um zu prüfen, ob er allgemein konvergiert, habe ich ihn in den Simulationen und Abschnitt 10 verwendet.

Der Beweis der Konvergenz von Algorithmus 1 wird folgende Resultate benötigen:

Hilfssatz 9.1 *Gegeben sei ein Datensatz* $\mathbf{Z}$ *mit den üblichen Bezeichnungen. Seien* $g^1 \geq g^0 \in \{0,1\}^n$ *mit* $\exists i \in \{1,\ldots,n\} : g_i^1 > g_i^0$. *Sei* $g^+ := g^1 - g^0$. *Sei* $\mathbf{X}_0 := \mathbf{X}(g^0)$, $y_0 := y(g^0)$, $\beta_0 := \beta(\mathbf{Z}(g^0)) = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'y_0$ *und analog* $\mathbf{X}_1, y_1, \beta_1, \mathbf{X}_+, y_+$. *Weiter sei*

$$\mathbf{V}_0 := \mathbf{X}_+(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_+', \quad \mathbf{V}_1 := \mathbf{X}_+(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'.$$

$\mathbf{X}_0$ *und* $\mathbf{X}_1$ *sollen vollen Spaltenrang haben. Dann gilt mit*

$$M(g) := (y(g) - \mathbf{X}(g)\beta(g))'(y(g) - \mathbf{X}(g)\beta(g)):$$
$$(\mathbf{I}_{n(g^+)} + \mathbf{V}_0)^{-1} = \mathbf{I}_{n(g^+)} - \mathbf{V}_1, \tag{9.2}$$
$$M(g^1) = M(g^0) + (y_+ - \mathbf{X}_+\beta_0)'(\mathbf{I}_{n(g^+)} + \mathbf{V}_0)^{-1}(y_+ - \mathbf{X}_+\beta_0), \tag{9.3}$$
$$M(g^0) = M(g^1) - (y_+ - \mathbf{X}_+\beta_1)'(\mathbf{I}_{n(g^+)} - \mathbf{V}_1)^{-1}(y_+ - \mathbf{X}_+\beta_1). \tag{9.4}$$

Beweis: Der Beweis von (9.2) und (9.3) findet sich in Plackett (1950) (Absatz 8 bzw. 10). Auf analoge Weise zu (9.3) läßt sich (9.4) beweisen: Sei $\mathbf{I} := \mathbf{I}_{n(g^+)}$. Es gilt

$$\mathbf{X}_+'y_+ = \mathbf{X}_1'y_1 - \mathbf{X}_0'y_0, \quad \mathbf{X}_+'\mathbf{X}_+ = \mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_0'\mathbf{X}_0.$$

Damit folgt

$$\mathbf{X}_+'(y_+ - \mathbf{X}_+\beta_0) = \mathbf{X}_1'y_1 - \mathbf{X}_0'y_0 - (\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_0'\mathbf{X}_0)\beta_0 = (\mathbf{X}_1'\mathbf{X}_1)(\beta_1 - \beta_0) \Rightarrow$$
$$\Rightarrow \mathbf{X}_+(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'(y_+ - \mathbf{X}_+\beta_0) = \mathbf{X}_+(\beta_1 - \beta_0) \Rightarrow$$

(Subtraktion von $(y_+ - \mathbf{X}_+\beta_0)$, Multiplikation mit -1)

$$\Rightarrow (\mathbf{I} - \mathbf{V}_1)(y_+ - \mathbf{X}_+\beta_0) = (y_+ - \mathbf{X}_+\beta_1) \Rightarrow \tag{9.5}$$
$$\Rightarrow (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'y_+ - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'\mathbf{X}_+\beta_0 =$$
$$= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'(\mathbf{I} - \mathbf{V}_1)^{-1}(y_+ - \mathbf{X}_+\beta_1). \tag{9.6}$$

Nun ist

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'y_+ = \beta_1 - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_0'y_0 \text{ und}$$
$$\mathbf{X}_+'\mathbf{X}_+\beta_0 = [(\mathbf{X}_1'\mathbf{X}_1) - (\mathbf{X}_0'\mathbf{X}_0)](\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'y_0.$$

Eingesetzt in die linke Seite aus (9.6) also:

$$\beta_1 - \beta_0 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_+'(\mathbf{I} - \mathbf{V}_1)^{-1}(y_+ - \mathbf{X}_+\beta_1). \tag{9.7}$$

Weiter ist $\mathbf{X}_1'\mathbf{X}_1\beta_1 = \mathbf{X}_1'y_1$, so daß

$$\begin{aligned}(y_1 - \mathbf{X}_1\beta_0)'(y_1 - \mathbf{X}_1\beta_0) &= (y_1 - \mathbf{X}_1\beta_1)'(y_1 - \mathbf{X}_1\beta_1) + (\beta_0 - \beta_1)'\mathbf{X}_1'\mathbf{X}_1(\beta_0 - \beta_1) = \\ &= M(g^1) + (y_+ - \mathbf{X}_+\beta_1)'(\mathbf{I} - \mathbf{V}_1)^{-1}\mathbf{V}_1(\mathbf{I} - \mathbf{V}_1)^{-1}(y_+ - \mathbf{X}_+\beta_1)\end{aligned} \tag{9.8}$$

wegen (9.7). (9.5) bringt

$$(y_+ - \mathbf{X}_+\beta_0)'(y_+ - \mathbf{X}_+\beta_0) = (y_+ - \mathbf{X}_+\beta_1)'(\mathbf{I} - \mathbf{V}_1)^{-1}(\mathbf{I} - \mathbf{V}_1)^{-1}(y_+ - \mathbf{X}_+\beta_1).$$

Zusammen mit (9.8) - die folgende Gleichung gilt nach Definition von $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_+, y_0, y_1, y_+$ - ergibt sich

$$\begin{aligned}(y_0 - \mathbf{X}_0'\beta_0)'(y_0 - \mathbf{X}_0'\beta_0) &= (y_1 - \mathbf{X}_1'\beta_0)'(y_1 - \mathbf{X}_1'\beta_0) - (y_+ - \mathbf{X}_+'\beta_0)'(y_+ - \mathbf{X}_+'\beta_0) = \\ &= M(g^1) - (y_+ - \mathbf{X}_+\beta_1)'(\mathbf{I} - \mathbf{V}_1)^{-1}(y_+ - \mathbf{X}_+\beta_1),\end{aligned}$$

also (9.4).

Satz 9.2 (Konvergenz) *Sei $c > 1$. Wenn $(\mathbf{X}(g)'\mathbf{X}(g))^{-1}$ für alle $g \in \{0,1\}^n$ mit $n(g) > p+1$ existiert, dann erreicht Algorithmus 1 nach einer endlichen Zahl von Schritten einen KQ-FPCV.*

Beweis: Zur Notation: Es gelten die Bezeichnungen aus Algorithmus 1. Das bedeutet: k sei gerade, $g_i^k = 1 \Rightarrow g_i^{k+1} = 1$ und $g_i^{k+1} = 0 \Rightarrow g_i^{k+2} = 0$. Mit „Schritten" sind immer die Schritte des Algorithmus gemeint. Außerdem

$$\beta(g) := \beta(\mathbf{Z}(g)), \qquad \sigma^2(g) := \sigma^2(\mathbf{Z}(g)),$$
$$P_m := \prod_{i=p+1}^{m-1}\left[1\left(\frac{c-1}{i-p} < 1\right)\left(1 - \frac{c-1}{i-p}\right) + 1\left(\frac{c-1}{i-p} \geq 1\right)\frac{1}{c}\right],$$
$$g^+ := g^{k+1} - g^k, \qquad g^- := g^{k+1} - g^{k+2},$$
$$y_+ := y(g^+), \quad \mathbf{X}_+ := \mathbf{X}(g^+), \quad y_- := y(g^-), \quad \mathbf{X}_- := \mathbf{X}(g^-),$$
$$\mathbf{V}_+(g) := \mathbf{X}_+(\mathbf{X}(g)'\mathbf{X}(g))^{-1}\mathbf{X}_+', \qquad \mathbf{V}_-(g) := \mathbf{X}_-(\mathbf{X}(g)'\mathbf{X}(g))^{-1}\mathbf{X}_-'.$$

1. Abschnitt: Ich zeige zunächst für $m \geq -1$ durch vollständige Induktion $n(g^{m+1}) > p+1$. $n(g^0) > p+1$ gilt nach Definition des Algorithmus 1. Zu zeigen ist: $n(g^m) > p+1 \Rightarrow n(g^{m+1}) > p+1$.

Induktionsschritt, Fall 1: Sei zunächst $\sigma^2(g^k) = 0$, $m = k$, also $g_i^k = 1 \Rightarrow \left(y_i - x_i'\beta(g^k)\right)^2 = 0$ und außerdem $n(g^k) > p+1$. Dann gilt nach Schritt 2

$$g_i^{k+1} = 1 \Leftrightarrow \left(y_i - x_i'\beta(g^k)\right)^2 = 0,$$
$$\sigma^2(g^{k+1}) = 0, \quad n(g^{k+1}) \geq n(g^k) > p+1.$$

Ist nun $\sigma^2(g^{k+1}) = 0$, $m = k+1$ und $n(g^{k+1}) > p+1$, so gilt nach Schritt 4 $g^{k+2} = g^{k+1}$, also $n(g^{k+2}) > p+1$. Damit ist der Induktionsschritt bereits durchgeführt. Es gilt aber sogar weiter nach Schritt 2

$$g_i^{k+3} = 1 \Leftrightarrow \left(y_i - x_i'\beta(g^{k+1})\right)^2 = 0, \text{ und } n(g^{k+3}) \geq n(g^{k+1}) > p+1.$$

Damit ist g^{k+3} KQ-FPCV bzgl. $\mathbf{Z}$, denn Schritt 4 und 2 bewirken keine Änderungen mehr, $g^{k+5} = g^{k+4} = g^{k+3}$ und damit $f(g^{k+3}) = g^{k+3}$ gemäß Definition 8.6. Somit ist für $\exists m: \ \sigma^2(g^m) = 0$ bereits der ganze Satz gezeigt.

Induktionsschritt, Fall 2: Sei also im ganzen folgenden Beweis $\sigma^2(g^m) > 0$ für $m = k, k+1, k+2$. Zu zeigen ist wieder

$$n(g^m) > p+1 \Rightarrow n(g^{m+1}) > p+1. \tag{9.9}$$

Nach Schritt 2 gilt $n(g^{k+1}) \geq n(g^k)$, also ist nur $m = k+1$ von Interesse. Nach Definiton

$$\begin{aligned} |\{i: g_i^{k+1} = 1 \wedge (y_i - x_i'\beta(g^m)^2 > c\sigma^2(g^m)\}| = n(g^-) \Rightarrow \\ \Rightarrow \sigma^2(g^m) \geq \tfrac{n(g^-)c\sigma^2(g^m)}{n(g^m)-p-1} \Leftrightarrow \\ \Leftrightarrow n(g^-) \leq \tfrac{n(g^m)-p-1}{c}. \end{aligned} \tag{9.10}$$

Daraus folgt im Falle $n(g^m) \geq p+c+1$:

$$\begin{aligned} n(g^{m+1}) = n(g^m) - n(g^-) \geq \left(1 - \tfrac{1}{c}\right) n(g^m) + \tfrac{p+1}{c} \geq \\ \geq \left(1 - \tfrac{1}{c}\right)(p+c+1) + \tfrac{p+1}{c} = p+c > p+1. \end{aligned}$$

Außerdem gilt mit $n(g^-) \in I\!N$ und (9.10)

$$n(g^m) < p+c+1 \Rightarrow 1 > n(g^-) = 0 \Rightarrow n(g^{m+1}) = n(g^m).$$

Also $n(g^{m+1}) > p+1$, der Induktionsschritt ist durchgeführt.

2. Abschnitt: Nun wird gezeigt, daß

$$T: \{0,1\}^n \mapsto [0,\infty): \ g \mapsto P_{n(g)}\sigma^2(\mathbf{Z}(g)) \tag{9.11}$$

in Schritt 2 und 4 des Algorithmus streng monoton fällt, falls $n(g^+) > 0$ bzw. $n(g^-) > 0$. Im Falle $n(g^+) = 0$ ist natürlich $T(g^{k+1}) = T(g^k)$, $n(g^-) = 0 \Rightarrow T(g^{k+2}) = T(g^{k+1})$.

Da $|\{0,1\}^n| < \infty$, bedeutet das, daß der Algorithmus in einer endlichen Zahl von Schritten die Situation $n(g^+) = 0 \wedge n(g^-) = 0$, also $g = g^k = g^{k+1} = g^{k+2}$ erreicht. Damit ist g ist KQ-FPCI bzgl. $\mathbf{Z}$, weil $n(g^+) = 0, n(g^-) = 0 \Leftrightarrow g = f(g)$ gemäß Definition

8.6.

Zur Monotonie von T. Es ist zu zeigen, daß T in Schritt 2 und 4 des Algorithmus verkleinert wird:

$$n(g^+) > 0 \Rightarrow T(g^{k+1}) < T(g^k). \tag{9.12}$$
$$n(g^-) > 0 \Rightarrow T(g^{k+2}) < T(g^{k+1}), \tag{9.13}$$

Die linke Seite wird im folgenden jeweils vorausgesetzt. Nach Definition gilt nun

$$(y_+ - \mathbf{X}'_+\beta(g^k))'(y_+ - \mathbf{X}'_+\beta(g^k)) \leq n(g_+)c\sigma^2(g^k) \text{ und}$$
$$(y_- - \mathbf{X}'_-\beta(g^{k+1}))'(y_- - \mathbf{X}'_-\beta(g^{k+1})) > n(g_-)c\sigma^2(g^{k+1}).$$

$\mathbf{V}_-(g^{k+2})$ und $\mathbf{V}_+(g^{k+1})$ sind offenbar positiv semidefinit und mit (9.2) gilt (setze in Hilfssatz 9.1 $g^0 = g^k$ bzw. g^{k+2} und $g^1 = g^{k+1}$)

$$(\mathbf{I}_{n(g^-)} - \mathbf{V}_-(g^{k+1}))^{-1} = \mathbf{I}_{n(g^-)} + \mathbf{V}_-(g^{k+2}),$$
$$(\mathbf{I}_{n(g^+)} + \mathbf{V}_+(g^k))^{-1} = \mathbf{I}_{n(g^+)} - \mathbf{V}_+(g^{k+1}).$$

Daher folgt auch

$$r_+^2 := (y_+ - \mathbf{X}_+\beta(g^k))'(\mathbf{I}_{n(g^+)} + \mathbf{V}_+(g^k))^{-1}(y_+ - \mathbf{X}_+\beta(g^k)) \leq n(g^+)c\sigma^2(g^k), \tag{9.14}$$
$$r_-^2 := (y_- - \mathbf{X}_-\beta(g^{k+1}))'(\mathbf{I}_{n(g^-)} - \mathbf{V}_-(g^{k+1}))^{-1}(y_- - \mathbf{X}_-\beta(g^{k+1})) > n(g^-)c\sigma^2(g^{k+1}). \tag{9.15}$$

Mit diesen Abschätzungen können (9.12) und (9.13) gezeigt werden. Dazu werden weiterhin folgende Abschätzungen benötigt:

$$\frac{n(g^{k+1})-p-1}{n(g^k)-p-1+n(g^+)c} > \frac{P_{n(g^{k+1})}}{P_{n(g^k)}}, \tag{9.16}$$
$$\frac{n(g^{k+1})-p-1-n(g^-)c}{n(g^{k+2})-p-1} \leq \frac{P_{n(g^{k+1})}}{P_{n(g^{k+2})}}. \tag{9.17}$$

Beweis von (9.16): Zur Vorbereitung zeige ich durch vollständige Induktion über m für $b > 0, m \in \mathbb{N}$ mit $a := c - 1$:

$$1 - \frac{am}{b+mc} > \left(1 - \frac{a}{b+m}\right)^{m_1}\left(\frac{1}{c}\right)^{m_2}$$
$$\forall m_1, m_2 \in \mathbb{N}_0 : m_1 + m_2 = m \tag{9.18}$$

unter der Voraussetzung

$$\frac{a}{b+m} < 1. \tag{9.19}$$

Für $m = m_1 = 1$ folgt (9.18) aus $c > 1$, für $m = m_2 = 1$ folgt es aus

$$1 - \frac{am}{b+mc} = \frac{b+m}{b+mc} > \frac{1}{c} \geq \left(\frac{1}{c}\right)^m. \tag{9.20}$$

Nun gelte (9.18) für $m = l - 1$ und beliebige $b > 0$ mit (9.19). Sei $l_1 + l_2 = l$. Wenn $l_2 = l$, dann gilt (9.18) für $m = l$ wegen (9.20). Sei daher nun $l_1 \geq 1$, $\frac{a}{b+l} < 1$. Damit ist

auch $\frac{a}{(b+1)+(l-1)} = \frac{a}{(b+1)+m} < 1$, so daß die Induktionsvoraussetzung angewendet werden kann:

$$\left(1-\tfrac{a}{b+l}\right)^{l_1}\left(\tfrac{1}{c}\right)^{l_2} = \left(1-\tfrac{a}{b+l}\right)\left(1-\tfrac{a}{(b+1)+(l-1)}\right)^{l_1-1}\left(\tfrac{1}{c}\right)^{l_2} <$$
$$< \left(1-\tfrac{a}{b+l}\right)\left(1-\tfrac{(l-1)a}{(b+1)+(l-1)c}\right) =$$
$$= 1-\tfrac{(l-1)a}{b+1+(l-1)c} - \tfrac{a(b+1)+a(l-1)c}{(b+l)(b+1+(l-1)c)} + \tfrac{(l-1)a^2}{(b+l)(b+1+(l-1)c)} =$$

(da $a^2 = ac - a$)

$$= 1-\tfrac{(l-1)a(b+l)+a(b+1)+(l-1)a}{(b+l)(b+1+(l-1)c)} =$$
$$= 1-\tfrac{(l-1)a}{b+lc} - \tfrac{(l-1)a(c-1)(b+l)+a(b+l)(b+lc)}{(b+1+(l-1)c)(b+lc)(b+l)} =$$
$$= 1-\tfrac{(l-1)a}{b+lc} - \tfrac{a(b+l)(b+lc+(l-1)(c-1))}{(b+1+(l-1)c)(b+lc)(b+l)} <$$

(da $b+lc \geq b+1+(l-1)c$)

$$< 1-\tfrac{(l-1)a}{b+lc} - \tfrac{a}{b+lc} = 1-l\tfrac{a}{b+lc} \qquad \text{q.e.d.}$$

Um (9.16) zu zeigen, sei nun zuerst $n(g^{k+1}) > p+c-1$. Es ist $n(g^{k+1}) > n(g^k) \geq p+1$, also unter Verwendung von (9.18) mit $m = n(g^+)$, $m_1 = n(g^{k+1}) - \max(n(g^k), \lceil p+c-1 \rceil)$, $b = n(g^k) - p - 1$, wobei (9.19) erfüllt ist, denn $\frac{a}{b+m} = \frac{c-1}{n(g^{k+1})-p-1} < 1$,

$$\frac{P_{n(g^{k+1})}}{P_n(g^k)} = \prod_{i=n(g^k)}^{n(g^{k+1})-1}\left(1\left(\frac{c-1}{i-p}<1\right)\left(1-\frac{c-1}{i-p}\right)+1\left(\frac{c-1}{i-p}\geq 1\right)\frac{1}{c}\right) =$$
$$= \prod_{\max(n(g^k),\lceil p+c-1\rceil)<i\leq n(g^{k+1})-1}\left(1-\frac{c-1}{i-p}\right)\prod_{i=n(g^k)}^{\lceil p+c-1\rceil}\frac{1}{c} \leq$$
$$\leq \left(1-\tfrac{c-1}{n(g^{k+1})-p-1}\right)^{m_1}\left(\tfrac{1}{c}\right)^{n(g^+)-m_1} <$$
$$< 1-\tfrac{(c-1)n(g^+)}{n(g^k)-p-1+n(g^+)c} = \tfrac{n(g^{k+1})-p-1}{n(g^k)-p-1+n(g^+)c}.$$

Also gilt in diesem Fall (9.16). Ist $n(g^{k+1}) \leq p+c-1$, so ist

$$\frac{P_{n(g^{k+1})}}{P_n(g^k)} = \left(\frac{1}{c}\right)^m < \frac{n(g^{k+1})-p-1}{n(g^k)-p-1+n(g^+)c}$$

wegen (9.20).

Beweis von (9.17): Falls $n(g^{k+2}) \leq p+c$, dann ist

$$\frac{n(g^{k+1})-p-1-n(g^-)c}{n(g^{k+2})-p-1} = 1-\frac{(c-1)n(g^-)}{n(g^{k+2})-p-1} \leq 0 < \frac{P_{n(g^{k+1})}}{P_{n(g^{k+2})}}.$$

Sei also $n(g^{k+1}) > n(g^{k+2}) > p+c$. Es folgt

$$\frac{P_{n(g^{k+1})}}{P_{n(g^{k+2})}} = \prod_{i=n(g^{k+2})}^{n(g^{k+1})-1}\left(1-\frac{c-1}{i-p}\right) \geq \left(1-\frac{c-1}{n(g^{k+2})-p-1}\right)^{n(g^-)}.$$

Nun gilt für $0 < b < 1, m \in I\!N$ durch vollständige Induktion $(1-b)^m \geq 1 - bm$: Für $m = 1$ gilt „=". Induktionsschritt:

$$(1-b)^m = (1-b)(1-b)^{m-1} \geq (1-b)(1-b(m-1)) = 1 - mb + (m-1)b^2 > 1 - mb.$$

Mit $b = \frac{c-1}{n(g^{k+2})-p-1}$, $m = n(g^-)$ folgt

$$\frac{P_{n(g^{k+1})}}{P_{n(g^{k+2})}} \geq 1 - n(g^-)\frac{c-1}{n(g^{k+2})-p-1}.$$

Beweis von (9.12): Setzt man $g^0 = g^k$ und $g^1 = g^{k+1}$ in Hilfssatz 9.1, so ist (9.3) gleichbedeutend zu

$$\sigma^2(g^{k+1}) = \tfrac{n(g^k)-p-1}{n(g^{k+1})-p-1}\sigma^2(g^k) + \\ + \tfrac{1}{n(g^{k+1})-p-1}(y_+ - \mathbf{X}_+\beta(g^k))'(\mathbf{I}_{n(g^+)} + \mathbf{V}_+(g^k))^{-1}(y_+ - \mathbf{X}_+\beta(g^k)).$$

Unter Verwendung von (9.14) und (9.16) erhält man

$$T(g^{k+1}) - T(g^k) = P_{n(g^{k+1})}\sigma^2(g^{k+1}) - P_{n(g^k)}\sigma^2(g^k) = \\ = \left(\tfrac{P_{n(g^{k+1})}(n(g^k)-p-1)}{n(g^{k+1})-p-1} - P_{n(g^k)}\right)\sigma^2(g^k) + \tfrac{P_{n(g^{k+1})}}{n(g^{k+1})-p-1}r_+^2 \leq \\ \leq \left(\tfrac{P_{n(g^{k+1})}(n(g^k)-p-1+n(g^+)c)}{n(g^{k+1})-p-1} - P_{n(g^k)}\right)\sigma^2(g^k) < 0.$$

Beweis von (9.13): Setzt man $g^0 = g^{k+2}$ und $g^1 = g^{k+1}$ in Hilfssatz 9.1, so ist (9.4) gleichbedeutend zu

$$\sigma^2(g^{k+2}) = \tfrac{n(g^{k+1})-p-1}{n(g^{k+2})-p-1}\sigma^2(g^{k+1}) - \\ - \tfrac{1}{n(g^{k+2})-p-1}(y_- - \mathbf{X}_-\beta(g^{k+1}))'(\mathbf{I}_{n(g^-)} - \mathbf{V}_-(g^{k+1}))^{-1}(y_- - \mathbf{X}_-\beta(g^{k+1})).$$

Unter Verwendung von (9.15) und (9.17) erhält man

$$T(g^{k+1}) - T(g^{k+2}) = P_{n(g^{k+1})}\sigma^2(g^{k+1}) - P_{n(g^{k+2})}\sigma^2(g^{k+2}) = \\ = \left(P_{n(g^{k+1})} - \tfrac{P_{n(g^{k+2})}(n(g^{k+1})-p-1)}{n(g^{k+2})-p-1}\right)\sigma^2(g^{k+1}) + \tfrac{P_{n(g^{k+2})}}{n(g^{k+2})-p-1}r_-^2 > \\ > \left(P_{n(g^{k+1})} - \tfrac{P_{n(g^{k+2})}(n(g^{k+1})-p-1-n(g^-)c)}{n(g^{k+2})-p-1}\right)\sigma^2(g^{k+1}) \geq 0.$$

Damit ist alles gezeigt.

10 Analyse von Beispieldatensätzen

Um die Arbeitsweise der Verfahren zu illustrieren, wird in diesem Abschnitt das Ergebnis der Analyse zweier Datensätze vorgestellt. Für die Berechnung der Fixpunktclusteranalyse wurde Algorithmus 2 aus Abschnitt 9 beginnend mit zufälligen Punktkonstellationen 140 mal mit $c = 10$ durchgerechnet. Eine genaue Beschreibung des Verfahrens findet sich in Abschnitt 15.1.1. Weiter wurde der Mischmodell-ML-Schätzer aus

Abschnitt 3.3 berechnet. Zur Ermittlung einer Approximation des globalen Maximums der Loglikelihood-Funktion pro vorgegebener Clusterzahl wurde das Maximum aus 20 Durchläufen des dort vorgestellten EM-Algorithmus ermittelt. Außerdem wurde der Fixed Partition-ML-Schätzer aus Abschnitt 3.4 berechnet, indem der dortige Algorithmus 50 mal pro vorgegebener Clusterzahl durchgeführt wurde. Die ML-Verfahren wurden mit vorgegebener Clusterzahl 1-5 durchgerechnet. Die Schätzung der Clusterzahl wurde wie in den Abschnitten 3.3 und 3.4 beschrieben durchgeführt. Eine genauere Beschreibung der Verfahren findet sich in Abschnitt 15, wobei in den Simulationen aber weniger Iterationen durchgeführt wurden als hier.

10.1 Telefondaten

Hierbei handelt es sich um den bereits in der Einleitung diskutierten Datensatz aus Rousseeuw und Leroy (1988). Zur Orientierung: Der von Rousseeuw und Leroy vorgeschlagene robuste Least Median of Squares-Schätzer, der nur die Jahre anpaßt, in denen die Telefonate gezählt wurden, ergibt $\hat{\beta}_{LMS} = (0.115, -5.610)$.

In 140 Durchläufen der Fixpunktclusteranalyse wurden 4 FPC gefunden:
115 mal wurde der Gesamtdatensatz als KQ-FPCV ($g^1 = (1,\ldots,1)$) gefunden. Die dazu gehörigen Schätzer sind die normalen KQ-Schätzer:

$$\beta(\mathbf{Z}(g^1)) = (0.504, -26.006),\ \sigma^2(\mathbf{Z}(g^1)) = 31.611.$$

22 mal wurde ein FPC gefunden, der die Punkte 1-14 und 22-24 enthält, d.h. die Jahre bis 1963 und ab 1971. Die Parameterschätzer:

$$\beta(\mathbf{Z}(g^2)) = (0.111, -5.260),\ \sigma^2(\mathbf{Z}(g^2)) = 0.0213.$$

Einmal wurde ein FPC gefunden, der die Punkte 1-13 und 22-24 enthält, d.h. dem zweiten FPC ohne den Wert für 1963 entspricht. Die Parameterschätzer:

$$\beta(\mathbf{Z}(g^3)) = (0.108, -5.164),\ \sigma^2(\mathbf{Z}(g^3)) = 0.0094.$$

Zweimal wurde ein FPC gefunden, der die Punkte 11 und 16-20, d.h. die Jahre 1960 und 1965-69 enthält. Die Parameterschätzer:

$$\beta(\mathbf{Z}(g^4)) = (2.150, -127.65),\ \sigma^2(\mathbf{Z}(g^4)) = 0.178.$$

Auffällig ist, daß der Gesamtdatensatz extrem häufig gefunden wurde. Das ist nach meiner Erfahrung fast immer der Fall: bei größeren Datensätzen fällt normalerweise der eine oder andere Punkt heraus. Dafür gibt es einen einfachen Grund: Wenn der Algorithmus mit Punkten gestartet wird, die nicht zum selben Cluster gehören, wird normalerweise eine recht hohe Störskala geschätzt. Die Schätzung für σ^2 ist schließlich, wie in Abschnitt 8.2 diskutiert, nicht robust. Dadurch wird der ganze oder ein großer Teil des Datensatzes nicht als Ausreißer klassifiziert. Wird dann die Ausreißeridentifikation auf der Basis des gesamten Datensatzes berechnet, ist es kaum möglich, Mitglieder unterschiedlicher vorhandener Cluster in Ausreißer und Nichtausreißer aufzuteilen. In vielen Datensätzen wird aufgrund mangelnder Robustheit der Parameterschätzer gar kein Ausreißer gefunden, so daß der Gesamtdatensatz ein KQ-FPC ist. Dieser Effekt muß bei der Interpretation des Ergebnisses einer Fixpunktclusteranalyse bekannt sein.

Des weiteren enthält Cluster 3 alle „guten" Daten im Sinne der robusten Statistik, d.h. die Daten, bei deren die Telefongespräche und nicht die Minuten gezählt wurden. Man kann dem Resultat entnehmen, daß das Jahr 1963, in welchem ja überwiegend „korrekt" gerechnet wurde, ein strittiger Punkt ist: Es taucht nur in Cluster 2 auf.

Der vierte Cluster enthält die Jahre, in denen die Minuten gezählt wurden, außer 1964. Dieses Jahr verursacht im wesentlichen den optischen Eindruck, es könne sich hier um einen nichtlinearen Zusammenhang handeln. Die dazugehörige Regressionsgerade geht fast genau durch den Punkt von 1960, der daher auch dazugehört. Dafür taucht das Jahr 1970 nur in dem Cluster auf, der dem Gesamtdatensatz entspricht.

Das Mischmodell-ML-Verfahren entscheidet sich mit Schwarz' Kriterium (3.1) für die Anzahl von vier Clustern. Das AIC (3.4) hätte sogar fünf Cluster geschätzt. Die vier Cluster sehen folgendermaßen aus:

Die Parameterschätzer für den ersten Cluster:

$$\hat{\beta}_1 = (0.108, -5.162),\ \hat{\sigma}_1^2 = 0.0086.$$

Die geschätzten Zugehörigkeitswahrscheinlichkeiten $\hat{\epsilon}_{i,1}$ für die Punkte i = 1-9, 11 und 13 sowie 22-24 sind fast 1, $\hat{\epsilon}_{10,1} = 0.602$. Für die restlichen Punkte (d.h. auch für $i = 12$, das Jahr 1961) gilt $\hat{\epsilon}_{i,1} \approx 0$.

Die Parameterschätzer für den zweiten Cluster:

$$\hat{\beta}_2 = (1.860, -107.14),\ \hat{\sigma}_2^2 = 0.00002.$$

Die geschätzten Zugehörigkeitswahrscheinlichkeiten $\hat{\epsilon}_{i,2}$ für die Punkte i = 15 und 20 (d.h. 1964 und 1969) sind 1, alle anderen fast 0.

Die Parameterschätzer für den dritten Cluster:

$$\hat{\beta}_3 = (0.312, -17.536),\ \hat{\sigma}_3^2 = 0.00002.$$

Die geschätzten Zugehörigkeitswahrscheinlichkeiten $\hat{\epsilon}_{i,3}$ für die Punkte i = 12, 14 und 21 sind fast 1, alle anderen fast 0.

Die Parameterschätzer für den vierten Cluster:

$$\hat{\beta}_4 = (1.889, -110.45),\ \hat{\sigma}_4^2 = 0.023.$$

Die geschätzten Zugehörigkeitswahrscheinlichkeiten $\hat{\epsilon}_{i,4}$ für die Punkte $i = 16 - 19$ (d.h. 1965-1968) sind 1, alle anderen fast 0 bis auf $\hat{\epsilon}_{10,4} = 0.398$.

Offenbar wird erkannt, daß die Mehrheit der Daten gut zusammenpaßt. Der entsprechende Cluster 1 läßt aber das Jahr 1961 heraus. Da die restlichen Werte nicht gut von einer gemeinsamen Gerade angepaßt werden können, werden sie in winzige Cluster aufgeteilt.

Beim Fixed Partition-ML-Verfahren wird mit dem modifizierten BIC (3.11) auf zwei Cluster entschieden. Cluster 1 enthält die Punkte 1-13 und 22-24, die Parameterschätzer entsprechen dem dritten FPC. Cluster 2 enthält die restlichen Punkte 14-21. Die Parameterschätzer:

$$\hat{\beta}_2 = (0.963, -51.487),\ \hat{\sigma}_2^2 = 43.177.$$

Der erste Cluster enthält also alle „guten Daten". Im zweiten Cluster werden die restlichen Punkte undifferenziert zusammengefaßt. Das führt zu einer schlechten Anpassung, gemessen durch $\hat{\sigma}_2^2$.

10.2 Artifizieller Datensatz

Der in Abbildung 8 gezeigte Datensatz wurde folgendermaßen erzeugt: Punkte 1-50 wurden unabhängig generiert entsprechend

$$\mathcal{L}(x) = \mathcal{N}_{(0,1)} \otimes \delta_1, \quad y = x'\beta_1 + e, \quad \beta_1 = (1, 0),$$

wobei $\mathcal{L}(e) = \mathcal{N}_{(0,0.01)}$ hier und auch für die anderen Punkte gilt. Die Punkte 51-98 wurden unabhängig generiert entsprechend

$$\mathcal{L}(x) = \mathcal{N}_{(0,1)} \otimes \delta_1, \quad y = x'\beta_2 + e, \quad \beta_2 = (-1, 0).$$

Die Punkte 99 und 100 wurden unabhängig generiert entsprechend

$$\mathcal{L}(x) = \mathcal{N}_{(5,1)} \otimes \delta_1, \quad y = e,$$

d.h. $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.01$. Der Datensatz soll ein Beispiel für ein deutliches Muster linearer Cluster mit Ausreißern sein. In diesem Datensatz wurden acht Fixpunktcluster

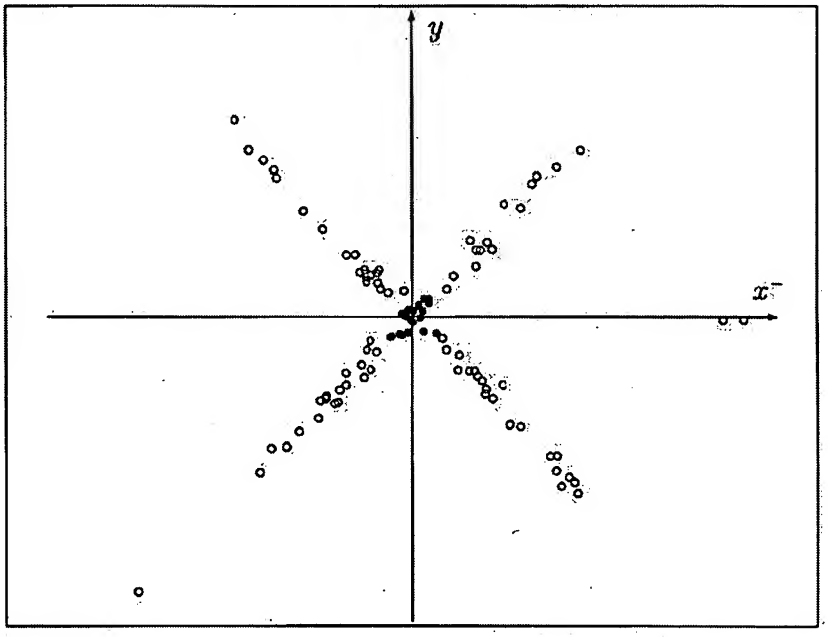


Abbildung 8: Artifizieller Datensatz

gefunden. Cluster 1 wurde 60 mal gefunden und ist wieder der gesamte Datensatz, die Parameterschätzer sind

$$\beta(\mathrm{Z}(g^1)) = (-0.058, -0.135), \ \sigma^2(\mathrm{Z}(g^1)) = 1.163.$$

Cluster 2 wurde 48 mal gefunden und entspricht dem gesamten Datensatz bis auf Punkt 14. Das ist der Punkt links unten mit dem minimalen y. Die Parameterschätzer sind

$$\beta(\mathrm{Z}(g^2)) = (-0.147, -0.090), \ \sigma^2(\mathrm{Z}(g^2)) = 1.044.$$

Cluster 3 wurde 12 mal gefunden und enthält 64 Punkte, nämlich die Punkte 51-98 und 16 weitere Punkte aus 1-50. In Abbildung 8 ist der Schnitt der Cluster 3 und 5, der alle diese Punkte enthält, voll ausgefüllt. Die Parameterschätzer sind

$$\beta(\mathbf{Z}(g^3)) = (-1.003, -0.007), \ \sigma^2(\mathbf{Z}(g^3)) = 0.028.$$

Cluster 4 wurde achtmal gefunden und enthält 62 Punkte, nämlich die Punkte 51-98 und 14 weitere Punkte aus 1-50. Er ist Teilmenge von Cluster 3. Die Parameterschätzer sind

$$\beta(\mathbf{Z}(g^4)) = (-1.009, 0.010), \ \sigma^2(\mathbf{Z}(g^4)) = 0.021.$$

Cluster 5 wurde sechsmal gefunden und enthält 53 Punkte, nämlich die Punkte 1-50 und drei weitere Punkte aus 51-98. Diese Punkte sind in Abbildung 8 ebenfalls voll ausgefüllt. Die Parameterschätzer sind

$$\beta(\mathbf{Z}(g^5)) = (0.998, -0.001), \ \sigma^2(\mathbf{Z}(g^5)) = 0.013.$$

Cluster 6 wurde viermal gefunden und enthält 55 Punkte, nämlich die 53 Punkte aus Cluster 5 und zwei weitere Punkte aus 51-98. Die Parameterschätzer sind

$$\beta(\mathbf{Z}(g^6)) = (0.998, 0.014), \ \sigma^2(\mathbf{Z}(g^6)) = 0.018.$$

Cluster 7 wurde einmal gefunden und enthält 63 Punkte, nämlich die 62 Punkte aus Cluster 4 und einen weiteren Punkt aus 1-50. Die Parameterschätzer sind

$$\beta(\mathbf{Z}(g^7)) = (-1.006, 0.002), \ \sigma^2(\mathbf{Z}(g^7)) = 0.025.$$

Cluster 8 wurde einmal gefunden und enthält 76 Punkte, nämlich die 64 Punkte aus Cluster 3 und zwölf weitere Punkte aus 1-50. Die Parameterschätzer sind

$$\beta(\mathbf{Z}(g^8)) = (-0.854, -0.120), \ \sigma^2(\mathbf{Z}(g^8)) = 0.243.$$

Neben dem bereits erwähnten Phänomen, daß am häufigsten der Gesamtdatensatz gefunden wird, fällt hier eine weitere Besonderheit ins Auge. Die Analyse bringt viel mehr Cluster, als deutlich im Datensatz zu sehen sind. Das ist im wesentlichen darauf zurückzuführen, daß einige Cluster in mehreren „Varianten" gefunden werden. Die Ausgabe aller Cluster kann zwar sehr unübersichtlich sein, vereinfacht sich jedoch, wenn man sieht, daß einige Cluster fast identisch sind, nämlich

- Cluster 1 und 2 (Gesamtdatensatz),
- Cluster 3, 4 und 7 und mit gutem Willen 8 (zweiter Modellcluster)
- Cluster 5 und 6 (erster Modellcluster).

Um sich diese Übersicht zu verschaffen, kann die Ausgabe der Cluster um eine Tabelle ergänzt werden, in der man sehen kann, wie häufig sie gefunden wurden und wieviele Punkte in ihrer jeweiligen Schnittmenge liegen. In diesem Beispiel sieht das folgendermaßen aus:

Cluster	gefunden	$n(g)$	1	2	3	4	5	6	7	8
1	60	100	100	99	64	62	53	55	63	76
2	48	99	99	99	64	62	52	54	63	76
3	12	64	64	64	64	62	19	21	63	64
4	8	62	62	62	62	62	17	19	62	62
5	6	53	53	52	19	17	53	53	18	31
6	4	55	55	54	21	19	53	55	20	33
7	1	63	63	63	63	62	18	20	63	63
8	1	76	76	76	64	62	31	33	63	76

Im Falle einer viel zu unübersichtlichen Ausgabe könnte man mit dieser Tabelle - interpretiert als Ähnlichkeitstabelle - noch eine angemessene Clusteranalyse rechnen. Wenn man an einer Parameterschätzung interessiert ist, kann man dafür die Cluster wählen, die aus einer Gruppe von ähnlichen Clustern am häufigsten gefunden wurden, d.h. hier Cluster 3 und 5.

Die Masse an Clustern ist Folge davon, daß sich die Cluster überschneiden können. Das hat den Vorteil, daß man detaillierte Informationen über die einzelnen Punkte bekommt. Zum Beispiel ist zu sehen, welche Punkte sich von beiden Modellen gut anpassen lassen, sofern man die Ausgabe so gedeutet hat, daß es im wesentlichen zwei Modelle gibt. Weiterhin ist zu sehen, daß die Ausreißer 99 und 100 nur in den Gesamtdatensatz-Clustern sind. Der Unterschied zwischen den Clustern 1 und 2 ist Punkt 14, der in gewissem Sinne auch als Ausreißer interpretiert werden kann. Andererseits sorgt die Überschneidung zwischen den Clustern dafür, daß die Cluster jeweils Punkte enthalten, die zu unterschiedlichen Modellkomponenten gehören. Das sorgt zumindest im Falle der zweiten Modellkomponente für eine deutliche Überschätzung von σ_2^2. Wie gut σ_2^2 in diesem Fall geschätzt werden könnte, zeigen die ML-Verfahren. Der zweite Modellcluster wird allerdings von allen Verfahren größer eingeschätzt als der erste.

Das Mischmodell-ML-Verfahren schätzt mit Schwarz' Kriterium wie mit dem AIC die Zahl von drei Clustern.

Cluster 1 hat einen Anteil von $\hat{\epsilon}_1 = 0.450$ am Gesamtdatensatz. Die Parameterschätzungen sind

$$\hat{\beta}_1 = (1.000, 0.002),\ \hat{\sigma}_1^2 = 0.010.$$

Cluster 2 hat einen Anteil von $\hat{\epsilon}_2 = 0.508$. Die Parameterschätzungen sind

$$\hat{\beta}_2 = (-1.014, 0.004),\ \hat{\sigma}_2^2 = 0.010.$$

Cluster 3 hat einen Anteil von $\hat{\epsilon}_3 = 0.043$. Die Parameterschätzungen sind

$$\hat{\beta}_3 = (0.047, -0.233),\ \hat{\sigma}_3^2 = 0.001.$$

Die geschätzten Wahrscheinlichkeiten für Punkt 99 und 100, zu diesem Cluster zu gehören, sind 1. Außerdem ist für Punkt 13 noch $\hat{\epsilon}_{13,3} = 0.536$. Einige weitere Punkte haben eine geschätzte Wahrscheinlichkeit zwischen 0.2 und 0.3. $\hat{\sigma}_3^2$ entspricht der vorgegebenen unteren Schranke für die Störvarianz (siehe Abschnitt 3.3). Diese Schranke hat möglicherweise großen Einfluß darauf, wie der kleinste Cluster konkret aussieht. In diesem Fall besteht er ja im wesentlichen aus zwei Punkten.

Da es nur zwei Ausreißer gibt, können sie durch eine gemeinsame Gerade angepaßt werden. Dadurch macht die drei-Cluster-Lösung einen vernünftigen Eindruck. Mehr Ausreißer würden vermutlich zu mehr Clustern führen und eventuell weitere Punkte aus den anderen Clustern herausschlagen.

Das Fixed Partition-ML-Verfahren schätzt mit dem modifizierten BIC ebenfalls drei Cluster.

Cluster 1 enthält die Punkte 1-50 außer 1, 5, 10, 17, 19 und 39. Die Parameterschätzungen sind

$$\hat{\beta}_1 = (1.000, 0.004), \ \hat{\sigma}_1^2 = 0.011.$$

Cluster 2 enthält die Punkte 51-98 außer 71 sowie die Punkte 1, 5, 17 und 39. Die Parameterschätzungen sind

$$\hat{\beta}_2 = (-1.014, 0.001), \ \hat{\sigma}_2^2 = 0.010.$$

Cluster 3 enthält die Punkte 10, 19, 71, 99 und 100. Die Parameterschätzungen sind

$$\hat{\beta}_2 = (0.042, -0.214), \ \hat{\sigma}_2^2 = 0.00002.$$

Um zu verhindern, daß die Likelihood degeneriert (siehe Abschnitt 3.4), war eine Mindestclustergröße von 4 vorgegeben. In diesem Fall wurde also eine vernünftige drei-Cluster-Lösung berechnet, weil es unter den Punkten 1-98 noch drei Punkte gab, die mit den Ausreißern etwa auf einer Gerade lagen. Hätte es keinen solchen Punkt gegeben, wäre die Behandlung der Ausreißer ein Problem gewesen. Aber schon hier läßt sich anhand der Ausgabe des Verfahrens nicht feststellen, daß die Punkte 99 und 100 von anderer Qualität sind als die Punkte 10, 19 und 71. Bei der Analyse der Residuen aller Punkte von allen Clustern würde das allerdings auffallen.

Artifizieller Datensatz								
Nr.	x	y	Nr.	x	y	Nr.	x	y
1	0.0964	-0.0131	35	0.1388	0.2008	69	0.5639	-0.4743
2	-0.9160	-1.0491	36	0.7766	0.7998	70	-0.1060	0.0361
3	-0.0497	-0.1925	37	0.1225	0.0593	71	0.2939	-0.2000
4	-0.5063	-0.6487	38	0.1989	0.1386	72	-0.4233	0.5105
5	0.0079	-0.0646	39	-0.1119	0.0390	73	1.7037	-1.6969
6	0.9691	0.8027	40	-0.5963	-0.7475	74	-0.6447	0.5246
7	0.7004	0.9089	41	0.9081	0.8876	75	1.1023	-0.8375
8	-1.1565	-1.2449	42	0.8253	0.7955	76	-0.5729	0.5102
9	-1.0611	-0.9764	43	2.0853	2.0017	77	1.7780	-1.8734
10	-0.1455	-0.2166	44	0.4945	0.4804	78	0.6911	-0.6678
11	-0.8228	-0.8411	45	-0.5506	-0.4101	79	-0.7054	0.7365
12	1.5433	1.6932	46	-0.2484	-0.2385	80	1.8391	-2.0637
13	-0.5082	-0.2972	47	-1.7297	-1.6009	81	0.7523	-0.6678
14	-3.3766	-3.3364	48	-0.9566	-1.0670	82	0.4088	-0.4093
15	-0.4290	-0.4364	49	-1.8666	-1.8987	83	-0.5914	0.5553
16	-0.8207	-0.6935	50	1.4846	1.6034	84	-1.1026	1.0496
17	-0.0717	0.0057	51	1.2096	-1.3342	85	2.0437	-2.1493
18	0.0134	0.0552	52	-0.0380	0.0752	86	0.8472	-0.7858
19	-0.1126	-0.2247	53	-2.1778	2.3731	87	-1.8206	1.8763
20	1.1250	1.3527	54	0.8943	-0.8931	88	-0.5681	0.4663
21	-1.3932	-1.4073	55	1.9362	-1.9498	89	-1.6988	1.7641
22	0.7729	0.6003	56	0.9768	-1.0006	90	1.3401	-1.3516
23	-1.1323	-1.0299	57	-0.8168	0.7334	91	0.1363	-0.1805
24	-0.0076	-0.0550	58	2.0057	-2.0173	92	-0.4923	0.4916
25	0.4113	0.3168	59	-0.0992	0.2998	93	-1.9983	2.0064
26	0.0751	0.1278	60	1.2102	-1.3147	94	-0.2817	0.2740
27	1.7855	1.7926	61	-0.3961	0.5508	95	-0.3784	0.3201
28	-0.2611	-0.2501	62	-1.3404	1.2689	96	-0.5576	0.4091
29	0.1992	0.2021	63	1.7832	-1.6952	97	-0.4135	0.3932
30	-1.5458	-1.5881	64	-0.0934	0.2912	98	-1.6657	1.6726
31	-0.6301	-0.5925	65	0.5465	-0.6624	99	4.0980	-0.0430
32	-0.8951	-0.9023	66	0.7906	-0.7325	100	3.8469	-0.0498
33	1.3421	1.3094	67	0.8860	-0.9463			
34	-1.0661	-1.0041	68	0.3643	-0.2649			

Teil III
Fixpunktclusterindikatoren in speziellen Modellen

11 Hilfsresultate

Um die Übersicht darüber zu gewährleisten, welche Resultate worauf aufbauen, habe ich diesen Abschnitt an den Anfang von Teil III gestellt. Die wesentlichen Resultate finden sich erst in den Abschnitten 12 und 13. Für das Verständnis der grundlegenden Ideen, die dort verwendet werden, werden aus diesem Abschnitt nur die Bezeichnungen zu Beginn des Unterabschnitts 11.2 benötigt.

11.1 Eigenschaften der Fixpunktcluster-Parameterfunktion

Für die Rechnung mit den Funktionalen, die in Definition 8.3 und Bemerkung 8.4 vorkommen, werden folgende Hilfssätze benötigt:

Hilfssatz 11.1 *Sei R ein Maß auf $(\mathbb{R}^{p+1} \times \mathbb{R}, \mathcal{B}^{p+2})$ mit*

$$\int y^2 dR(x,y) < \infty, \quad \int \|x\|^2 dR(x,y) < \infty. \tag{11.1}$$

Dann gilt für $i = 1, \ldots, p+1$:

$$\frac{\partial}{\partial t_i} \int (y - x't)^2 dR(x,y) = -2 \int x_i (y - x't) dR(x,y). \tag{11.2}$$

Ist darüberhinaus

$$\int xx' dR(x,y) \text{ invertierbar,} \tag{11.3}$$

so ist

$$\arg\min_{\beta} \int (y - x'\beta)^2 dR(x,y) = \left[\int xx' dR(x,y)\right]^{-1} \int xy dR(x,y)$$

existent und eindeutig.

Beweis: Nach Korollar 16.3 aus Bauer (1990) darf für eine Funktion

$$\phi(t) := \int f(t,\omega) d\mu(\omega), \; x \in U \text{ offen } \subseteq \mathbb{R}^d$$

partielle Differentiation nach t_i und Integration vertauscht werden, wenn eine μ-integrierbare Funktion j auf Ω (dem Maßraum, auf dem μ definiert ist) existiert, so daß

$$\left|\frac{\partial}{\partial t_i} f(t,\omega)\right| \leq j(\omega) \; \forall (t,\omega) \in U \times \Omega. \tag{11.4}$$

Außerdem muß $\omega \mapsto f(t,\omega)$ für jedes t μ–integrierbar und $t \mapsto f(t,\omega)$ für beliebiges ω nach jedem t_i partiell differenzierbar sein.

Für $f_0(t,x,y) := (y - x't)^2$ gilt in einer Umgebung eines beliebigen $\theta \in I\!R^{p+1}$, d.h. für $\|t-\theta\| < \epsilon$ mit $\epsilon > 0$:

$$\left|\frac{\partial}{\partial t_i} f_0(t,x,y)\right| = |-2x_i(y - x't)| \le \|x\|^2 + y^2 + 2\|x\|^2(\|\theta\| + \epsilon).$$

Letzteres ist nach Voraussetzung (11.1) R-integrierbar. f_0 ist nach jedem t_i partiell differenzierbar und nach (11.1) für gegebenes t integrierbar. Also folgt (11.2). Sei

$$\beta(R) := \left[\int xx'dR(x,y)\right]^{-1} \int xydR(x,y), \qquad h(t) := \int (y - x't)^2 dR(x,y).$$

$\beta(R)$ existiert nach Voraussetzung (11.3). Soll θ lokales Minimum von h sein, muß für $i = 1,\ldots,p+1$

$$\frac{\partial}{\partial t_i}\int (y - x't)^2 dR(x,y)\bigg|_{t=\theta} = -2\int x_i(y - x'\theta)dR(x,y) \stackrel{!}{=} 0,$$

d.h. zusammengefaßt für $i = 1,\ldots,p+1$

$$\int xydR(x,y) \stackrel{!}{=} \int xx'dR(x,y)\theta$$

gelten. Das ist äquivalent zu $\theta = \beta(R)$. Für $f_i(t,x,y) := x_i(y - x't)$ gilt

$$\left|\frac{\partial}{\partial t_j} f_i(t,x,y)\right| = -x_ix_j.$$

Letzteres ist mit $\|x\|^2$ R-integrierbar, so daß die Hessesche Matrix H_h von h wie folgt aussieht:

$$H_h(t) = \left(\frac{\partial^2}{\partial t_i \partial t_j}\int (y - x't)^2 dR(x,y)\right)_{i,j=1,\ldots,p+1} = 2\int xx'dR(x,y).$$

$H_h(t)$ kann keine negativen Eigenwerte haben, da $\int v'xx'vdR(x,y) \ge 0$ für beliebiges $v \in I\!R^{p+1}$. Weiter ist $H_h(t)$ nach Voraussetzung (11.3) invertierbar und damit überall positiv definit. Also ist $\beta(R)$ eindeutiges lokales Minimum von h und h ist konvex. Daher ist $\beta(R)$ auch eindeutiges globales Minimum.

Hilfssatz 11.2 *Es gelten die Bezeichnungen von Bemerkung 8.4. Sei $M \subseteq I\!R^{p+1} \times I\!R^+$. Weiter sei*

$$\int y^2 dQ(x,y) < \infty, \ \int \|x\|^2 dQ(x,y) < \infty, \tag{11.5}$$

$$Q\{(y - x'\theta)^2 = cs^2\} = 0 \quad \forall(\theta,s^2) \in M, \tag{11.6}$$

$$Q\{(y - x'\theta)^2 < cs^2\} > 0 \quad \forall(\theta,s^2) \in M, \tag{11.7}$$

$$\int xx'g_{\theta,s^2}(x,y)dQ(x,y) \text{ invertierbar } \forall(\theta,s^2) \in M. \tag{11.8}$$

Dann ist die Einschränkung von f (definiert gemäß (8.9)) auf M stetig.

Beweis: Sei zunächst h eine Abbildung von $\mathbb{R}^{p+2}$ nach $\mathbb{R}$, $a < b \in \mathbb{R}, d \in \mathbb{R}^{p+2}$. Ich zeige

$$l(a,b,d) := \int h(z)1(a \le z'd \le b)dQ(z) \text{ ist stetig in } (a,b,d) \tag{11.9}$$

unter den Voraussetzungen

$$Q\{z : a = z'd\} = Q\{z : b = z'd\} = 0, \quad \int |h(z)|dQ(z) < \infty.$$

Beweis von (11.9): Es gilt wegen des Satzes für majorisierte Konvergenz

$$\lim_{n\to\infty}(a_n,b_n,d_n) = (a,b,d) \Rightarrow \lim_{n\to\infty} l(a_n,b_n,d_n) = l(a,b,d), \text{ denn}$$
$$\forall n: \quad l(a_n,b_n,d_n) \le \int |h(z)|dQ(z) < \infty \text{ und}$$
$$\lim_{n\to\infty} h(z)1(a_n \le z'd_n \le b_n) = h(z)1(a \le z'd \le b) \quad \forall z \text{ mit } a \ne z'd \ne b, \text{ also}$$
$$Q\{z : h(z)1(a_n \le z'd_n \le b_n) \to h(z)1(a \le z'd \le b)\} = 1.$$

Also gilt (11.9).

Sei nun (θ, s^2) immer aus M und

$$\beta(\theta,s^2) := \beta(g_{\theta,s^2},Q), \quad \sigma^2(\theta,s^2) := \sigma^2(g_{\theta,s^2},Q).$$

Hilfssatz 11.1 bringt mit $dR = g_{\theta,s^2}dQ$, wobei mit (11.5) und (11.8) die Voraussetzungen (11.1) und (11.3) erfüllt sind:

$$\beta(\theta,s^2) = \left(\int xx'g_{\theta,s^2}(x,y)dQ(x,y)\right)^{-1} \int xyg_{\theta,s^2}(x,y)dQ(x,y). \tag{11.10}$$

Sei nun für $i,j = 1,\ldots,p+1$

$$l^0_{ij}(a,b,d) := \int x_ix_j1[a \le (x',y)d \le b]dQ(x,y),$$
$$l_{ij}(\theta,s^2) := \int x_ix_jg_{\theta,s^2}(x,y)dQ(x,y) = l^0_{ij}[-\sqrt{c}s, \sqrt{c}s, (-\theta',1)'].$$

l_{ij} ist stetig in (θ,s^2), da wegen $|x_ix_j| \le \|x\|^2$, (11.5) und (11.6) die Voraussetzungen für (11.9) erfüllt sind und damit l^0_{ij} stetig in a,b,d ist. Damit ist aber auch

$$\left(\int xx'g_{\theta,s^2}(x,y)dQ(x,y)\right)^{-1}$$

komponentenweise stetig in (θ,s^2), da die Komponenten einer inversen Matrix im Falle der in (11.8) vorausgesetzten Existenz Quotienten von Summen von Produkten der Komponenten der zu invertierenden Matrix, also stetige Funktionen dieser Komponenten (d.h. der $l_{ij}(\theta,s^2)$) sind. Weiter liefert (11.9) für $i = 1,\ldots,p+1$ die Stetigkeit in (θ,s^2) von $\int x_iyg_{\theta,s^2}(x,y)dQ(x,y)$, da $\int |x_iy|dQ(x,y) < \infty$ mit (11.5), so daß mit (11.6) wieder die Voraussetzungen erfüllt sind. Zusammengesetzt haben wir nun die Stetigkeit von β nach (11.10).

Weiterhin ist auch σ^2 definiert in (8.7) stetig: Der Nenner $\int g_{\theta,s^2}(x,y)dQ(x,y)$ ist in (11.7) als positiv vorausgesetzt und mit (11.9) stetig in (θ,s^2). Nun fehlt noch die Stetigkeit des Zählers. Es gilt

$$m(\beta_0,x,y) = (y - x'\beta_0)^2 = y^2 - 2yx'\beta_0 + (x'\beta_0)^2 \le y^2 + \|\beta_0\|(y^2 + \|x^2\|) + \|\beta_0\|^2\|x\|^2,$$

also $\int m(\beta_0, x, y)dQ(x,y) < \infty$ mit(11.5). Also ist mit (11.9) für gegebenes β_0

$$m_1(\beta_0, \theta, s^2) := \int m(\beta_0, x, y) g_{\theta,s^2} dQ(x,y)$$

stetig in (θ, s^2). Weiter ist m_1 auch stetig in β_0: Für $\beta_n \to_{n\to\infty} \beta_0$ gilt:

$$|m_1(\beta_n, \theta, s^2) - m_1(\beta_0, \theta, s^2)| = |\int [(y - x'\beta_n)^2 - (y - x'\beta_0)^2]\, g_{\theta,s^2}(x,y)dQ(x,y)| =$$
$$= |\int 2yx'(\beta_0 - \beta_n) g_{\theta,s^2}(x,y)dQ(x,y) + \int x'Bx g_{\theta,s^2}(x,y)dQ(x,y)| \leq$$

(wobei $B := (b_{ij})_{i,j=1,\ldots,p+1} := \beta_n\beta_n' - \beta_0\beta_0'$)

$$\leq \int (\|\theta\|\|x\| + \sqrt{c}s)\|x\|\|\beta_0 - \beta_n\| dQ(x,y) + \int \|x\|^2 (p+1)^2 \max_{i,j=1,\ldots,p+1} |b_{ij}| dQ(x,y) \to 0$$

wegen $\max|b_{ij}| \to 0$ und Voraussetzung (11.5). Der Zähler aus (8.7) ist nun gleich $m_1(\beta(\theta, s^2), \theta, s^2)$, also, nachdem die Stetigkeit von β bereits gezeigt ist, ebenfalls stetig in (θ, s^2). Damit ist auch σ^2 und somit f stetig auf M.

11.2 Abgeschnittene Normalverteilungen

Abgeschnittene Normalverteilungen tauchen immer auf, wenn ein normalverteilter Term mit einer der für die Fixpunktclusteranalyse benötigten Indikatorfunktionen g multipliziert wird. Für die Ergebnisse in den folgenden Abschnitten werden einige Hilfssätze über das Verhalten von Erwartungswert und Varianz abgeschnittener Normalverteilungen benötigt. Es gelten für den Rest der Arbeit folgende Bezeichnungen ($u \in \mathbb{R}, s > 0$):

$$E(u,s) := \frac{\int y 1((y-u)^2 \leq s^2) d\mathcal{N}(y)}{\int 1((y-u)^2 \leq s^2) d\mathcal{N}(y)} = \frac{\varphi(u-s) - \varphi(u+s)}{\Phi(u+s) - \Phi(u-s)},$$
$$V(u,s) := \frac{\int (y - E(u,s))^2 1((y-u)^2 \leq s^2) d\mathcal{N}(y)}{\int 1((y-u)^2 \leq s^2) d\mathcal{N}(y)},$$
$$E_+(u,s) := \frac{\varphi(u-s) + \varphi(u+s)}{\Phi(u+s) - \Phi(u-s)}.$$

Offenbar gilt

$$E(-u,s) = -E(u,s), \quad V(-u,s) = V(u,s), \quad E_+(-u,s) = E_+(u,s). \tag{11.11}$$

Hilfssatz 11.3

$$\begin{aligned}
\tfrac{\partial}{\partial s}E(u,s) &= (E(u,s)-u)E(u,s)+sE_+(u,s),\\
\tfrac{\partial}{\partial t}E(u,s) &= (u-E(u,s))E_+(u,s)-sE(u,s),\\
\tfrac{\partial}{\partial u}E_+(u,s) &= (E(u,s)-u)E_+(u,s)+sE(u,s),\\
\tfrac{\partial}{\partial s}E_+(u,s) &= uE(u,s)-(s+E_+(u,s))E_+(u,s),\\
V(u,s) &= 1+\tfrac{(u-s)\varphi(u-s)-(u+s)\varphi(u+s)}{\Phi(u+s)-\Phi(u-s)}-E(u,s)^2 = \qquad (11.12)\\
&= 1+[u-E(u,s)]E(u,s)-sE_+(u,s), \qquad (11.13)
\end{aligned}$$

$$\begin{aligned}
\tfrac{\partial}{\partial u}V(u,s) &= E(u,s)[1-u^2-s^2+3uE(u,s)-3sE_+(u,s)-2E^2(u,s)]+2usE_+(u,s),\\
\tfrac{\partial}{\partial s}V(u,s) &= 2sE(u,s)[E(u,s)-u]+uE_+(u,s)[u-E(u,s)]+(s^2-1)E_+(u,s)+\\
&\quad +E_+(u,s)[sE_+(u,s)+2E(u,s)^2-2uE(u,s)].
\end{aligned}$$

Beweis: Die ersten vier Gleichungen ergeben sich mit Hilfe von

$$\begin{aligned}
\tfrac{\partial}{\partial u}[\varphi(u-s)-\varphi(u+s)] &= -(u-s)\varphi(u-s)+(u+s)\varphi(u+s) =\\
&= [\Phi(u+s)-\Phi(u-s)][sE_+(u,s)-uE(u,s)],\\
\tfrac{\partial}{\partial s}[\varphi(u-s)-\varphi(u+s)] &= (u-s)\varphi(u-s)+(u+s)\varphi(u+s) =\\
&= [\Phi(u+s)-\Phi(u-s)][uE_+(u,s)-sE(u,s)],\\
\tfrac{\partial}{\partial u}[\varphi(u-s)+\varphi(u+s)] &= -(u-s)\varphi(u-s)-(u+s)\varphi(u+s) =\\
&= [\Phi(u+s)-\Phi(u-s)][sE(u,s)-uE_+(u,s)],\\
\tfrac{\partial}{\partial s}[\varphi(u-s)+\varphi(u+s)] &= [\Phi(u+s)-\Phi(u-s)][uE(u,s)-sE_+(u,s)],\\
\tfrac{\partial}{\partial u}[\Phi(u+s)-\Phi(u-s)] &= \varphi(u+s)-\varphi(u-s) = -[\Phi(u+s)-\Phi(u-s)]E(u,s),\\
\tfrac{\partial}{\partial s}[\Phi(u+s)-\Phi(u-s)] &= \varphi(u+s)+\varphi(u-s) = [\Phi(u+s)-\Phi(u-s)]E_+(u,s).
\end{aligned}$$

Beweis von (11.12): Für den Zähler gilt

$$\begin{aligned}
z &:= \int_{u-s}^{u+s}(y-E(u,s))^2\varphi(y)dy =\\
&= \int_{u-s}^{u+s}(y-E(u,s))y\varphi(y)dy-E(u,s)\int_{u-s}^{u+s}(y-E(u,s))\varphi(y)dy.
\end{aligned}$$

Das letzte Integral ist 0 nach Definition von $E(u,s)$. Mit partieller Integration:

$$\begin{aligned}
z &= \int_{u-s}^{u+s}\varphi(y)dy-[(y-E(u,s))\varphi(y)]_{u-s}^{u+s} =\\
&= \Phi(u+s)-\Phi(u-s)+(u-s)\varphi(u-s)-(u+s)\varphi(u+s)+\\
&\quad +E(u,s)\varphi(u+s)-E(u,s)\varphi(u-s).
\end{aligned}$$

Der Nenner ist $\int_{u-s}^{u+s}\varphi(y)dy = \Phi(u+s)-\Phi(u-s)$, womit sich (11.12) ergibt. (11.13) folgt aus (11.12).

Die letzten beiden Gleichungen folgen mit (11.13) und den ersten vier Gleichungen, wobei

$$\begin{aligned}
\tfrac{\partial}{\partial u}V(u,s) &= (u-E(u,s))[(E(u,s)-u)E(u,s)+sE_+(u,s)]+\\
&+E(u,s)[1-(E(u,s)-u)E(u,s)-sE_+(u,s)]-\\
&-s[(E(u,s)-u)E_+(u,s)+sE(u,s)],\\
\tfrac{\partial}{\partial s}V(u,s) &= (u-E(u,s))\,[(u-E(u,s))E_+(u,s)-sE(u,s)]+\\
&+E(u,s)\,[sE(u,s)-(u-E(u,s))E_+(u,s)]-\\
&-E_+(u,s)-s\,[u(E(u,s)-(s+E_+(u,s))E_+(u,s)]=\\
&=2sE(u,s)^2+u^2E_+(u,s)+s^2E_+(u,s)+sE_+(u,s)^2+\\
&+2E(u,s)^2E_+(u,s)-2suE(u,s)-3uE_+(u,s)E(u,s)-E_+(u,s).
\end{aligned}$$

Hilfssatz 11.4 *Für $s>0$ ist $f'(u)<0$ mit $f(u):=E(u,s)-u$ und*

$$\lim_{u\to\infty} f(u) = -s. \tag{11.14}$$

Beweis:

$$f(u)=E(u,s)-u=\frac{\int_{u-s}^{u+s}t\varphi(t)dt}{\int_{u-s}^{u+s}\varphi(t)dt}-u=\frac{\int_{-s}^{s}t\varphi(t+u)dt}{\int_{-s}^{s}\varphi(t+u)dt}\Rightarrow \tag{11.15}$$

$$\Rightarrow f'(u)=-\frac{\int_{-s}^{s}(t^2+tu)\varphi(t+u)dt}{\int_{-s}^{s}\varphi(t+u)dt}+\frac{\int_{-s}^{s}(t+u)\varphi(t+u)dt\int_{-s}^{s}t\varphi(t+u)dt}{[\int_{-s}^{s}\varphi(t+u)dt]^2}=$$

$$=-\frac{\int_{-s}^{s}t^2\varphi(t+u)dt}{\int_{-s}^{s}\varphi(t+u)dt}+\left(\frac{\int_{-s}^{s}t\varphi(t+u)dt}{\int_{-s}^{s}\varphi(t+u)dt}\right)^2<0.$$

Beweis von (11.14): Sei X_u eine Zufallsvariable mit transformierter abgeschnittener Normalverteilung, d.h. mit Riemann-Dichte

$$\varphi_u(x)=\frac{\varphi(x+u)}{\int_{-s}^{s}\varphi(t+u)dt}1(|x|\le s).$$

Dann ist $f(u)=EX_u$ wegen (11.15). Es gilt

$$\forall\epsilon>0:\lim_{u\to\infty}\frac{P(|X_u-(-s)|\ge\epsilon)}{P(|X_u-(-s)|<\frac{\epsilon}{2})}=$$

$$=\lim_{u\to\infty}\frac{\int_{-s+\epsilon}^{s}\varphi(x+u)dx}{\int_{-s}^{-s+\frac{\epsilon}{2}}\varphi(x+u)dx}\le\lim_{u\to\infty}\frac{(2s-\epsilon)\varphi(-s+\epsilon+u)}{\frac{\epsilon}{2}\varphi(-s+\frac{\epsilon}{2}+u)}=0,$$

$$\text{da }\lim_{t\to\infty}\frac{\varphi(t+\frac{\epsilon}{2})}{\varphi(t)}=0.$$

Also konvergiert X_u stochastisch gegen $-s$. Da $|X_u|\le s$, gilt $EX_u^2<s^2\ \forall u$ und damit nach dem Korollar zu Satz 25.12 aus Billingsley (1986) auch $\lim_{u\to\infty}EX_u=-s$.

Hilfssatz 11.5 *Für $s > 0$ gilt*

$$E_+(0,s) > 0 = E(0,s),$$
$$u - s < E(u,s) < u + s,$$
$$u > 0 \Rightarrow 0 < E(u,s) < u, \quad u < 0 \Rightarrow u < E(u,s) < 0,$$
$$u \geq 0 \Rightarrow E(u,s) < E_+(u,s) < E(u,s) + \tfrac{1}{s},$$
$$\tfrac{\partial}{\partial s} E_+(u,s) < 0.$$

Beweis: Die erste Gleichung und $E_+(u,s) > |E(u,s)| \geq E(u,s)$ sind nach Definition klar.

$$u - s < E(u,s) = E_{\mathcal{N}}(X | u - s \leq X \leq u + s) < u + s,$$
$$u > 0 \Rightarrow \varphi(u-s) > \varphi(u+s) \Rightarrow E(u,s) > 0.$$

$E(u,s) < u$ folgt für $u > 0$ aus $E(0,s) = 0$ und Hilfssatz 11.4, die entsprechende Aussage für $u < 0$ gilt wegen (11.11).

$$\forall u \geq 0: \ \varphi(u+s) = \inf\{\varphi(t) : u - s \leq t \leq u + s\} \Rightarrow$$
$$\Rightarrow E(u,s) + \tfrac{1}{s} > E(u,s) + \tfrac{2s\varphi(u+s)}{s(\Phi(u+s)-\Phi(u-s))} = E_+(u,s) > E(u,s).$$

Die vierte Gleichung aus Hilfssatz 11.3 und $E_+(-u,s) = E_+(u,s)$ liefern $\frac{\partial}{\partial s} E_+(u,s) < 0$, denn mit dem bisher Gezeigten gilt

$$E_+(u,s) > E(u,s) \geq 0, \quad u \geq 0 \Rightarrow E_+(u,s) + s > E(u,s) + s > u \geq 0.$$

Hilfssatz 11.6 *$E(u,s)$ ist streng monoton fallend in $s > 0$ falls $u > 0$ bzw. streng monoton steigend falls $u < 0$. $E_+(u,s)$ ist für $s > 0$ streng monoton steigend in $u > 0$ und streng monoton fallend in $u < 0$.*

Beweis: Sei $s_1 > s_2 > 0$, $u > 0$. Dann gilt wegen (11.15) folgende Äquivalenz:

$$E(u,s_1) - E(u,s_2) < 0 \Leftrightarrow$$
$$\Leftrightarrow \frac{\int_{-s_1}^{s_1} v\varphi(v+u)dv}{\int_{-s_1}^{s_1} \varphi(v+u)dv} - \frac{\int_{-s_2}^{s_2} t\varphi(t+u)dt}{\int_{-s_2}^{s_2} \varphi(t+u)dt} < 0 \Leftrightarrow$$
$$\Leftrightarrow D(s_1,s_2) := \int_{-s_1}^{s_1} dv \int_{-s_2}^{s_2} dt[v\varphi(v+u)\varphi(t+u) - t\varphi(v+u)\varphi(t+u)] < 0.$$
$$D(s_1,s_2) = \int_{-s_1}^{-s_2} dv \int_{-s_2}^{s_2} dt(v-t)\varphi(v+u)\varphi(t+u) +$$
$$+ \int_{s_2}^{s_1} dv \int_{-s_2}^{s_2} dt(v-t)\varphi(v+u)\varphi(t+u) +$$
$$+ \int_{-s_2}^{s_2} dv \int_{-s_2}^{s_2} dt(v-t)\varphi(v+u)\varphi(t+u) =$$

(Der letzte Summand ist aus Symmetriegründen 0, im ersten Summanden Substitution $v \to -v$.)

$$= \int_{s_2}^{s_1} dv \int_{-s_2}^{s_2} dt[(-v-t)\varphi(-v+u) + (v-t)\varphi(v+u)]\varphi(t+u) =$$

(Aufteilung des dt-Integrals und Substitution $t \to -t$ im negativen Teil.)

$$= \int_{s_2}^{s_1} dv \int_0^{s_2} dt[v(\varphi(v+u) - \varphi(-v+u))(\varphi(t+u) + \varphi(-t+u)) - \\ -t(\varphi(t+u) - \varphi(-t+u))(\varphi(v+u) + \varphi(-v+u))].$$

Ich zeige nun, daß der letzte Term kleiner als 0 ist, da der Integrand für $v > t \geq 0$ immer kleiner als 0 ist. Das ist äquivalent zu $E(u,s_1) - E(u,s_2) < 0$, was zu zeigen war. Es ist für $s > 0$

$$\frac{\partial}{\partial s}\left(\frac{\varphi(s+u)-\varphi(-s+u)}{\varphi(s+u)+\varphi(-s+u)}\right) = \\ = \frac{-[\varphi(s+u)+\varphi(-s+u)][(s-u)\varphi(s+u)+(-s+u)\varphi(-s+u)]+[\varphi(s+u)-\varphi(-s+u)][(s+u)\varphi(s+u)-(-s+u)\varphi(-s+u)]}{[\varphi(s+u)+\varphi(-s+u)]^2} = \\ = \frac{u[(\varphi(s+u)-\varphi(-s+u))^2-(\varphi(s+u)+\varphi(-s+u))^2]}{(\varphi(s+u)+\varphi(-s+u))^2} < 0,$$

und daher für $v > t \geq 0$:

$$[\varphi(v+u) - \varphi(-v+u)][\varphi(t+u) + \varphi(-t+u)] < \\ < [\varphi(v+u) + \varphi(-v+u)][\varphi(t+u) - \varphi(-t+u)] \leq 0.$$

Das Verhalten für $u < 0$ folgt aus (11.11). Weiter gilt mit Hilfssatz 11.3:

$$\frac{\partial}{\partial u}E_+(u,s) = -\frac{\partial}{\partial s}E(u,s),$$

also ist $E_+(u,s)$ für $s > 0$ streng monoton steigend in $u > 0$ und streng monoton fallend in $u < 0$.

Hilfssatz 11.7 *Für $s > 0$ gilt*

$$1 - \frac{\partial}{\partial u}E(u,s) = V(u,s) < 1,$$

also auch $\frac{\partial}{\partial u}E(u,s) > 0$. Weiter ist $V(0,s)$ streng monoton steigend in s.

Beweis: Hilfssatz 11.3 bringt

$$1 - \frac{\partial}{\partial u}E(u,s) = 1 + (u - E(u,s))E(u,s) - sE_+(u,s) = V(u,s).$$

Das ist kleiner als 1, da $s > u - E(u,s)$ und $E_+(u,s) > E(u,s)$ wegen Hilfssatz 11.5. Für $s_1 > s_2 > 0$ ist

$$V(0,s_1) = \frac{\int y^2 1(y^2 \leq s_1^2) d\mathcal{N}(y)}{\int 1(y^2 \leq s_1^2) d\mathcal{N}(y)} = \\ = \frac{\int 1(s_2^2 < y^2 \leq s_1^2) d\mathcal{N}(y)}{\int 1(y^2 \leq s_1^2) d\mathcal{N}(y)} \left(\frac{\int y^2 1(s_2^2 < y^2 \leq s_1^2) d\mathcal{N}(y)}{\int 1(s_2^2 < y^2 \leq s_1^2) d\mathcal{N}(y)}\right) + \frac{\int 1(y^2 \leq s_2^2) d\mathcal{N}(y)}{\int 1(y^2 \leq s_1^2) d\mathcal{N}(y)} V(0,s_2).$$

Das ist größer als $V(0,s_2)$, da

$$\frac{\int y^2 1(s_2^2 < y^2 \le s_1^2) d\mathcal{N}(y)}{\int 1(s_2^2 < y^2 \le s_1^2) d\mathcal{N}(y)} > s_2^2 \ge E_{\mathcal{N}}(Y^2 | Y^2 \le s_2^2) = V(0,s_2).$$

Damit ist alles gezeigt.

Hilfssatz 11.8 *$|u| < 0.63$, $s \ge 1 \Rightarrow V(u,s)$ ist streng monoton steigend in s.*

Beweis: Sei wegen (11.11) ohne Einschränkung $u \ge 0$. Es seien für $r \in \mathbb{R}$, $s \in [1,\infty)$

$$b_0(r,s) := \int_{u-s}^{u+s} (y-r)^2 \varphi(y) dy,\ r_1(s) := E(u,s),$$
$$r_2(s) := s,\ r(s) := (r_1(s), r_2(s)),\ b(s) := b_0(r(s)),$$
$$P(s) := \int_{u-s}^{u+s} \varphi(y) dy.$$

Alle diese Funktionen sind offenbar in allen Komponenten stetig und differenzierbar. Außerdem ist $V(u,s) = \frac{b(s)}{P(s)}$. Dann gilt:

$$b'(s) = \lim_{h\to 0} \frac{1}{h} \left[\int_{u-s-h}^{u+s+h} (y - E(u,s+h))^2 \varphi(y) dy - \int_{u-s}^{u+s} (y - E(u,s))^2 \varphi(y) dy \right] =$$
$$= \lim_{h\to 0} \frac{P(s+h)}{h} \left[\frac{b(s+h)}{P(s+h)} - \frac{P(s)}{P(s+h)} V(u,s) \right] =$$
$$= \lim_{h\to 0} P(s+h) \lim_{h\to 0} \frac{1}{h} \left[\frac{b(s+h)}{P(s+h)} - V(u,s) + \frac{P(s+h) - P(s)}{P(s+h)} V(u,s) \right] =$$
$$P(s) \frac{\partial}{\partial s} V(u,s) + \lim_{h\to 0} \frac{1}{h} [P(s+h) - P(s)] V(u,s), \text{ also}$$
$$\frac{\partial}{\partial s} V(u,s) = \frac{1}{P(s)} [b'(s) - P'(s) V(u,s)]. \qquad (11.16)$$

Einerseits ist nun $P'(s) = \varphi(u+s) + \varphi(u-s)$. Andererseits sind

$$\frac{\partial}{\partial r} b_0(r,s) = -2 \int_{u-s}^{u+s} (y-r) \varphi(y) dy,$$
$$\frac{\partial}{\partial s} b_0(r,s) = (u+s-r)^2 \varphi(u+s) + (u-s-r)^2 \varphi(u-s).$$

Damit gilt

$$b'(s) = \text{grad } b_0(r_1(s), r_2(s))^t \begin{pmatrix} r_1'(s) \\ r_2'(s) \end{pmatrix} =$$
$$= (u+s-E(u,s))^2 \varphi(u+s) + (u-s-E(u,s))^2 \varphi(u-s) -$$
$$-2 r_1'(s) \int_{u-s}^{u+s} (y - E(u,s)) \varphi(y) dy =$$
$$= (u+s-E(u,s))^2 \varphi(u+s) + (u-s-E(u,s))^2 \varphi(u-s),$$

denn das Integral ist 0 nach Definition von $E(u,s)$. Eingesetzt in (11.16) ergibt sich also

$$P(s)\frac{\partial}{\partial s}V(u,s) = \Big([s+u-E(u,s)]^2 - V(u,s)\Big)\varphi(u+s) +$$

$$+\Big([s-(u-E(u,s))]^2 - V(u,s)\Big)\varphi(u-s).$$

Der Hilfssatz ist bewiesen, wenn für $s \geq 1, u < 0.63$: $\frac{\partial}{\partial s}V(u,s) > 0$, also

$$W(u,s) :=$$

$$\Big([s+u-E(u,s)]^2 - V(u,s)\Big)\frac{\varphi(u+s)}{\varphi(u-s)} + \Big([s-(u-E(u,s))]^2 - V(u,s)\Big) > 0. \quad (11.17)$$

Sei zunächst $s \geq 1.63$. Hilfssatz 11.5 liefert $u \geq E(u,s) \geq 0$, also

$$s - u - E(u,s) \geq s - u - E(u,s) \geq s - u \geq 1.$$

Nach Hilfssatz 11.7 ist $V(u,s) < 1$. Also gilt (11.17).

Im Folgenden sei $1 \leq s < 1.63$. $V(u,s)$ ist die Varianz einer unimodalen Verteilung auf einem Träger $[u-s, u+s]$ der Länge $2s$. Nach Theorem 3 aus Jacobson (1969) ist die Varianz einer solchen Verteilung maximal $\frac{(\text{Trägerlänge})^2}{9}$. Daher $V(u,s) \leq \frac{4}{9}s^2$. Nach Hilfssatz 11.4 und Hilfssatz 11.6 ist $u - E(u,s)$ maximal, wenn u und s maximal gewählt werden, also

$$u - E(u,s) \leq 0.63 - E(0.63, 1.63) = 0.3757.$$

Weiter ist $s + u - E(u,s) \geq s \geq 1$. Zusammen:

$$W(u,s) \geq \left(1 - \tfrac{4}{9}\right)\frac{\varphi(u+s)}{\varphi(u-s)}s^2 + \left((s-0.3757)^2 - \tfrac{4}{9}s^2\right) =$$
$$= s^2\left[\left(1 - \tfrac{4}{9}\right)\frac{\varphi(u+s)}{\varphi(u-s)} + \left(\tfrac{(s-0.3757)^2}{s^2} - \tfrac{4}{9}\right)\right].$$

Mit $h(s) := \frac{(s-0.3757)^2}{s^2}$ ist $h'(s) = \frac{2(s-0.3757)0.3757}{s^3} > 0$, also

$$W(u,s) \geq s^2\left[\left(1 - \tfrac{4}{9}\right)\frac{\varphi(u+s)}{\varphi(u-s)} + \left((1-0.3757)^2 - \tfrac{4}{9}\right)\right] > 0 \Leftrightarrow$$
$$\Leftrightarrow W_0(u,s) := \tfrac{5}{9}\frac{\varphi(u+s)}{\varphi(u-s)} - 0.0547 > 0.$$

W_0 wird minimiert, wenn $\frac{\varphi(u+s)}{\varphi(u-s)}$ minimiert wird. Es sind mit $S(u,s) := \varphi(u+s)\varphi(u-s) > 0$, $s, u \geq 0$:

$$\frac{\partial}{\partial u}\frac{\varphi(u+s)}{\varphi(u-s)} = \frac{-(u+s)S(u,s) + (u-s)S(u,s)}{\varphi(u-s)^2} \leq 0,$$
$$\frac{\partial}{\partial s}\frac{\varphi(u+s)}{\varphi(u-s)} = \frac{-(u+s)S(u,s) - (u-s)S(u,s)}{\varphi(u-s)^2} \leq 0.$$

Also wird $W_0(u,s)$ minimiert, wenn u und s maximal gewählt werden, also $u = 0.63$, $s = 1.63$. Insgesamt folgt

$$W_0(u,s) \geq 0.5556\frac{0.031}{0.242} - 0.0547 = 0.0165 > 0,$$

womit der Hilfssatz gezeigt ist.

12 Fixpunktclusterindikatoren in homogenen Modellen

In diesem Abschnitt werden homogene Regressionsmodelle mit Regressionsparameter β_0 und Störskala σ_0^2 behandelt. Solche Verteilungen erzeugen normalerweise nicht mehrere anschauliche Cluster. Damit steht die folgende Theorie in Einklang: Im Falle eines normalverteilten Störterms zeige ich Existenz und Eindeutigkeit eines FPCI g mit $\beta(g,P)=\beta_0$ mit den Bezeichungen aus Definition 8.3. Dieser Satz ist ein zentrales Ergebnis, da er das Verhalten der KQ-FPCI in den Verteilungen aus $\mathcal{P}_0$ nach Definition 7.4 beschreibt. Diese Verteilungen bestimmen in der Definition der KQ-FPCI die Begriffe „Ausreißer" und „zusammengehörig". Ist der Störterm unimodal mit beschränktem Träger symmetrisch um 0 verteilt, wird noch die Existenz eines solchen FPCI gezeigt. Letztere Situation ist ein Beispiel für eine alternative Verteilungsklasse, die auch homogene Regressionsdatensätze erzeugt.

Satz 12.1 (Homogene Population mit normalverteiltem Störterm) *Sei* $c > 3$,

$$P(x,y) = \int 1(t \le x)\Phi_{0,\sigma_0^2}(y - t'\beta_0)dG(t)$$

wie in Modell 3 aus Abschnitt 2, wobei $(E_G(xx'))^{-1}$ *und* $E_G(\|x\|^2)$ *existieren sollen und* $\sigma_0^2 \ge 0$. *Dann existiert genau ein KQ-FPCI* g *bzgl.* P. *Für* g *gilt*

$$\beta(g,P) = \beta_0, \qquad \sigma^2(g,P) = k\sigma_0^2, \tag{12.1}$$

wobei $k > 0$ *die eindeutige Nullstelle ist von*

$$h(k) := 1 - k - \frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck}) - \Phi(-\sqrt{ck})}.$$

Beweis: Zunächst sei $\sigma_0^2 > 0$. Aufgrund der Äquivarianzeigenschaften von KQ-FPCI nach Bemerkung 8.5 sei ohne Beschränkung der Allgemeinheit $\beta_0 = 0, \sigma_0^2 = 1$, also insbesondere x und y stochastisch unabhängig. Der Beweis ist folgendermaßen gegliedert:

Schritt 1: Es existiert eine eindeutige positive Nullstelle von h.

Schritt 2: Für $g(x,y) = 1(y^2 \le cs^2)$ gilt: g ist KQ-FPCI bzgl. $P \Leftrightarrow s^2 = k$, wobei k positive Nullstelle von h ist. Für dieses g gilt (12.1).

Schritt 3: Für $\theta \ne 0$ ist $g(x,y) = 1\left((y - x'\theta)^2 \le cs^2\right)$ nicht FPCI bzgl. P. Da ein KQ-FPCI g nach Definiton 8.3 die Form $g(x,y) = 1\left((y - x'\theta)^2 \le cs^2\right)$ mit $s^2 \ge 0, \theta \in \mathbb{R}^{p+1}$ haben muß, ist damit alles gezeigt.

Schritt 4: Der Satz gilt auch für $\sigma_0^2 = 0$.

Beweis von Schritt 1: Sei $s > 0$. Dann gilt die Äquivalenz

$$h(s^2) = 0 \Leftrightarrow h_0(s) := (1 - s^2)[\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s)] - 2\sqrt{c}s\varphi(\sqrt{c}s) = 0.$$

Für $s \geq 1$ gilt $h_0(s) < 0$. Ich zeige, daß $h_0(s) > 0$ für $s > 0$ nahe genug an 0. Damit hat h eine positive Nullstelle s_0 wegen der Stetigkeit von h_0 und des Zwischenwertsatzes.
Es ist $h_0(0) = 0$ und für $s > 0$

$$\begin{aligned} h_0'(s) &= -2s[\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s)] + (1-s^2)2\sqrt{c}\varphi(\sqrt{c}s) + \\ &\quad + 2\sqrt{c^3}s^2\varphi(\sqrt{c}s) - 2\sqrt{c}\varphi(\sqrt{c}s) = \\ &= 2s[\sqrt{c}s(c-1)\varphi(\sqrt{c}s) - (\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s))] > \\ &> 2\sqrt{c}s^2[(c-1)\varphi(\sqrt{c}s) - 2\varphi(0)]. \end{aligned}$$

Letzteres ist größer als 0, falls $(c-1)\varphi(\sqrt{c}s) - 2\varphi(0) > 0$. Das gilt, wenn $s > 0$ klein genug ist, da $c - 1 > 2$ nach Voraussetzung des Satzes. Also gilt $h_0'(s) > 0$ und damit $h_0(s) > 0$ für s in einer positiven Nachbarschaft von 0.

Es wird nun die Eindeutigkeit der positiven Nullstelle von h gezeigt. Es gilt die Äquivalenz

$$h_0'(s) = 0 \Leftrightarrow h_1(s) := \sqrt{c}s(c-1)\varphi(\sqrt{c}s) - (\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s)) = 0,$$
$$\text{wobei } \lim_{s\to\infty} h_1(s) = -1.$$

Weiter ist $h_1(0) = 0$ und $h_1(s)$ hat für alle $s > 0$ dasselbe Vorzeichen wie $h_0'(s)$. Ich zeige nun: h_1 hat ein eindeutiges lokales Maximum s_2 mit $h_1(s_2) > 0$, für $s > s_2$ ist h_1 streng monoton fallend. Daraus ergibt sich, daß h_1 eine eindeutige Nullstelle s_1 hat, die damit auch eindeutige Nullstelle von h_0' und eindeutiges lokales Maximum von h_0 ist. Für $s > s_1$ muß also h_0 streng monoton fallend sein, so daß die Nullstelle s_0 von h_0 bzw. h eindeutig ist.
Es ist

$$\begin{aligned} h_1'(s) &= (c-1)\sqrt{c}(1-cs^2)\varphi(\sqrt{c}s) - 2\sqrt{c}\varphi(\sqrt{c}s) = \\ &= \sqrt{c}\varphi(\sqrt{c}s)[(c-1)(1-cs^2) - 2] \text{ und} \\ h_1'(0) &= \sqrt{c}\varphi(0)(c-1-2) > 0. \end{aligned}$$

Also gilt

$$h_1'(s) < 0 \Leftrightarrow (c-1)(1-cs^2) - 2 < 0.$$

$(c-1)(1-cs^2) - 2$ ist streng monoton fallend in $s > 0$, d.h. h_1' hat eine eindeutige Nullstelle s_2 und h_1 fällt streng monoton für $s > s_2$. Insgesamt ist nun Existenz und Eindeutigkeit der positiven Nullstelle von h gezeigt.

Beweis von Schritt 2: Sei $g(x,y) = 1(y^2 \leq cs^2)$ gegeben. Zu untersuchen sind (8.3), (8.4) und (8.5) aus Definition 8.3. Für $s^2 = 0$ ist $\int g(x,y)dP(x,y) = 0$, also gilt (8.3) nicht. Sei im Folgenden $s^2 > 0$. Jetzt folgt (8.3) aus $cs^2 > 0$. Mit $E_G(xx')$ ist auch

$$\int xx'g(x,y)dP(x,y) = \int 1(y^2 \leq cs^2)d\mathcal{N}(y) \int xx'dG(x)$$

invertierbar und $\int y^2 g(x,y)dP(x,y) = V(0,\sqrt{c}s)\int g(x,y)d\mathcal{N}(y) < 1$ nach Hilfssatz 11.7. Also sind für $dR = gdP$ die Voraussetzungen von Hilfssatz 11.1 erfüllt, wobei

aus der G–Intergrierbarkeit von $\|x\|^2$ auch die P– und R–Integrierbarkeit folgt. Damit gilt für g auch (8.4) und

$$\beta(g,P) = \arg\min_\beta \int\int (y - x'\beta)^2 g(x,y) d\mathcal{N}(y) dG(x) =$$

$$= \left[\int xx' g(x,y) dG(x)\right]^{-1} \int xyg(x,y) d\mathcal{N}(y) dG(x) = 0,$$

$$\text{da } \int xyg(x,y) d\mathcal{N}(y) dG(x) = \int y 1(y^2 < cs^2) d\mathcal{N}(y) \int x dG(x) = 0$$

wegen der Unabhängigkeit von x und y. Damit (und mit Hilfssatz 11.3)

$$\sigma^2(g,P) = \frac{\int y^2 1(y^2 \le cs^2) d\mathcal{N}(y)}{\int 1(y^2 \le cs^2) d\mathcal{N}(y)} = V(0, \sqrt{c}s) = 1 - \frac{2\sqrt{c}s\varphi(\sqrt{c}s)}{\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s)},$$

d.h. $\sigma^2(g,P) = s^2 \Leftrightarrow h(s^2) = 0$. Für $\beta(g,P) = 0$ ist (8.5) genau dann erfüllt (und damit g KQ-FPCI bzgl. P), wenn $\sigma^2(g,P) = s^2$.

Beweis von Schritt 3: Angenommen, es gäbe einen KQ-FPCI g mit

$$g(x,y) = 1\left((y - x'\theta)^2 \le cs^2\right),\ s^2 \ge 0, \theta \ne 0.$$

Der Fall $s^2 = 0$ scheidet wieder aus, da dann (8.3) verletzt wäre. Nach (8.5) wäre

$$\theta = \arg\min_\beta \int (y - x'\beta)^2 1\left((y - x'\theta)^2 \le cs^2\right) dP(x,y). \tag{12.2}$$

Sei für $t \in \mathbb{R}^{p+1}$

$$F_\theta(t) := \int\int (y - x't)^2 1\left((y - x'\theta)^2 \le cs^2\right) d\mathcal{N}(y) dG(x). \tag{12.3}$$

Mit $dR = g_{\theta,s^2} dP$, wobei $g_{\theta,s^2}(x,y) = 1((y - x'\theta)^2 \le cs^2)$, ist wieder Voraussetzung (11.1) von Hilfssatz 11.1 erfüllt. Also gilt mit (11.2):

$$v := \frac{\theta}{\|\theta\|}, \qquad \frac{\partial}{\partial v} F_\theta(t) := < v, \operatorname{grad} F_\theta(t) > =$$

$$= -2\sum_{i=1}^p v_i \int x_i \int (y - x't) 1\left((y - x'\theta)^2 \le cs^2\right) d\mathcal{N}(y) dG(x), \text{ also}$$

$$\frac{\partial}{\partial v} F_\theta(\theta) = -\frac{2}{\|\theta\|} \int x'\theta J(x'\theta) dG(x) \text{ mit}$$

$$J(u) := \int (y - u) 1\left((y - u)^2 \le c_0^2\right) d\mathcal{N}(y), \quad c_0 := \sqrt{cs^2}.$$

Es müßte $\frac{\partial}{\partial v} F_\theta(\theta) = 0$ sein, um (12.2) zu erfüllen. Wäre $uJ(u) < 0\ \forall u \ne 0$, so folgte $\frac{\partial}{\partial v} F_\theta(\theta) > 0$, da $G\{x'\theta \ne 0\} > 0$ für $\theta \ne 0$; anderenfalls wäre $\theta' E_G(xx')\theta = 0$ und $E_G(xx')$ nicht invertierbar. Ich zeige also nur noch $uJ(u) < 0$ für $u \ne 0$:

$$uJ(u) = \int u(y - u) 1(|y - u| \le c_0) \varphi(y) dy =$$

$$= \int u|y - u| 1(|y - u| \le c_0)[1(y > u) - 1(y < u)] \varphi(y) dy =$$

$$= \int u|t|1(|t| \le c_0)[1(t>0) - 1(t<0)]\varphi(t+u)dt =$$
$$= \int u|t|1(0 < t \le c_0)[\varphi(t+u) - \varphi(-t+u)]dt =$$
$$= \int |u||t|1(0 < t \le c_0)[\varphi(t+|u|) - \varphi(-t+|u|)]dt, \qquad (12.4)$$
$$\text{da für } u<0: \ \varphi(t+u) = \varphi(-t+|u|), \quad \varphi(-t+u) = \varphi(t+|u|).$$

φ ist symmetrisch um 0 und im Positiven streng monoton fallend, so daß

$$t>0, w>0 \Rightarrow wt[\varphi(t+w) - \varphi(-t+w)] < 0.$$

Aus (12.4) folgt $uJ(u) < 0$.

Beweis von Schritt 4: Sei nun $\sigma_0^2 = 0$, ohne Einschränkung $\beta_0 = 0$, d.h. $P\{y=0\} = 1$. Ich zeige:

$$g(x,y) = 1((y - x'\theta)^2 \le cs^2) \text{ ist KQ-FPCI } \Leftrightarrow (\theta, s^2) = (0,0). \qquad (12.5)$$

Für g aus (12.5) mit (θ, s^2) beliebig gilt für $\beta = 0$:

$$\int (y - x'\beta)^2 g(x,y)dP(x,y) = \int (y - x'\beta)^2 g(x,y)1(y^2 = 0)dP(x,y) = 0$$

und für $\beta \neq 0$:

$$\int (y - x'\beta)^2 g(x,y)dP(x,y) \ge 0.$$

Das heißt: Entweder $\beta(g,P) = \arg\min_\beta \int (y - x'\beta)^2 g(x,y)dP(x,y)$ nicht eindeutig, so daß (8.4) nicht erfüllt ist, oder $\beta(g,P) = 0$. Also muß $\theta = 0$ sein, um (8.5) zu erfüllen. Für $\theta = 0$ und beliebiges $s^2 \ge 0$ sind (8.3) und (8.4) erfüllt und es gilt $\sigma^2(g,P) = 0$, so daß g genau dann KQ-FPCI ist, wenn $s^2 = 0$.

Bemerkung 12.2 *Die Konstante k ist unabhängig von β_0, σ_0^2, G. Es ist also $\sigma_0^2(g,P) := \frac{\sigma^2(g,P)}{k} = \sigma_0^2$ für den eindeutigen KQ-FPCI g aus Satz 12.1. Für einen gegebenen Datensatz $\mathbf{Z}$ mit eindeutigem FPCV g ist also $\sigma_0^2(\mathbf{Z}(g)) := \frac{\sigma^2(\mathbf{Z}(g))}{k}$ Fisher-konsistenter Schätzer für σ_0^2. Für $c = 10$ ergibt sich zum Beispiel $k = 0.9795$.*

Bemerkung 12.3 *Satz 12.1 besagt, daß der einzige KQ-FPCI für $P \in \mathcal{P}_0$ die Form $g(x,y) = 1[(y - x'\beta_0)^2 \le ck\sigma_0^2]$ hat. Nach (8.2) ist das genau $1 - 1[\bullet \in A(\alpha_1, P)]$ mit $\alpha_1 = 1 - \chi_1^2(ck) = 0.0018$ für $c = 10$. Für KQ-FPCI im Regressionsfall und $\mathcal{P}_0$ ist also die Ausreißereigenschaft erfüllt (siehe Definition 7.5).*

Satz 12.4 (Homogene Population mit beschränktem Störträger) *Sei* $c > 3$,

$$P(x,y) = \int 1(t \le x) Q(y - t'\beta_0) dG(t),$$

wobei $(E_G(xx'))^{-1}$ *und* $E_G(\|x\|^2)$ *existieren sollen.* Q *habe einen beschränkten Träger* $\operatorname{supp} Q = [-m, m]$, $m \in \mathbb{R}^+$ *und eine beschränkte Riemann-Dichte* q. q *sei symmetrisch um 0, stetig in 0 und streng monoton fallend zwischen 0 und* m. *Dann existiert ein KQ-FPCI* g *bzgl.* P *mit*

$$\beta(g,P) = \beta_0, \qquad \sigma^2(g,P) = k \int y^2 dQ(y) = k \operatorname{Var}(Q), \tag{12.6}$$

wobei $1 \ge k > 0$ *Nullstelle ist von*

$$j(k) := \frac{\int y^2 1(y^2 \le ck) dQ(y)}{\int 1(y^2 \le ck) dQ(y)} - k.$$

g ist KQ-FPCI bzgl. $P \Leftrightarrow$ (12.6), so daß g genau dann eindeutig ist, wenn die positive Nullstelle k von j eindeutig ist.

Beweis: Es wird analog zum Beweis von Satz 12.1 vorgegangen: Aufgrund der Äquivarianzeigenschaften von KQ-FPCI nach Bemerkung 8.5 sei ohne Beschränkung der Allgemeinheit $\beta_0 = 0, E_Q(y^2) = \operatorname{Var}(Q) = 1$, also insbesondere x und y stochastisch unabhängig. Der Beweis ist folgendermaßen gegliedert:

Schritt 1: Es existiert eine positive Nullstelle $k \le 1$ von j.

Schritt 2: Für $g(x,y) = 1(y^2 \le cs^2)$ gilt: g ist KQ-FPCI bzgl. $P \Leftrightarrow s^2 = k$. Für solche g gilt (12.6).

Schritt 3: Für $\theta \ne 0$ ist $g(x,y) = 1((y - x'\theta)^2 \le cs^2)$ nicht FPCI bzgl. P.

Beweis von Schritt 1: j ist stetig für Argumente größer als 0, weil Riemann-Integrale stetige Funktionen ihrer Intervallgrenzen sind und

$$\int 1(y^2 \le ck) dQ(y) = \int_{-\sqrt{ck}}^{\sqrt{ck}} q(y) dy > 0$$

wegen $q(0) > 0$ und q stetig. Sei zunächst $s^2 > 1$. Es ist

$$j(s^2) = E_Q(y^2 | y^2 \le cs^2) - s^2 \le E_Q(y^2) - s^2 = 1 - s^2 < 0.$$

Alle Nullstellen von j sind also ≤ 1. Die Existenz einer positiven Nullstelle von j folgt also mit dem Zwischenwertsatz aus der Existenz von $s^2 > 0$ mit $j(s^2) \ge 0$. Angenommen, ein solches s^2 würde nicht existieren. Es gilt

$$\forall s > 0 : j(s^2) < 0 \Leftrightarrow$$
$$\Leftrightarrow \forall s > 0 : \frac{\int (y^2 - s^2) 1(y^2 \le cs^2) dQ(y)}{\int 1(y^2 \le cs^2) dQ(y)} < 0 \Leftrightarrow$$

$$\Leftrightarrow \forall s > 0 : \int (y^2 - s^2) 1(y^2 \leq cs^2) dQ(y) < 0 \Leftrightarrow$$
$$\Leftrightarrow \forall s > 0 : \int (y^2 - s^2) 1(s^2 < y^2 \leq cs^2) dQ(y) <$$
$$< \int (s^2 - y^2) 1(y^2 \leq s^2) dQ(y). \qquad (12.7)$$

Nun ist einerseits, da q zwischen 0 und m streng monoton fällt,

$$\int (y^2 - s^2) 1(s^2 < y^2 \leq cs^2) dQ(y) \geq$$
$$\geq \int (y^2 - s^2) 1(2s^2 < y^2 \leq cs^2) dQ(y) \geq s^2 (c-2) s^2 q(cs^2),$$

und andererseits

$$\int (s^2 - y^2) 1(y^2 \leq s^2) dQ(y) < (s^2)^2 q(0).$$

Zusammen mit der Stetigkeit von q in 0 gilt

$$(12.7) \Rightarrow \forall s > 0 : q(0) > (c-2) q(cs^2) \Rightarrow$$
$$\Rightarrow q(0) \geq \lim_{s \to 0} (c-2) q(cs^2) = (c-2) q(0).$$

Das ist ein Widerspruch zu $c - 2 > 1$ und $q(0) > 0$.

Beweis von Schritt 2: Sei $g(x,y) = 1(y^2 \leq cs^2)$ gegeben. Zu untersuchen sind (8.3), (8.4) und (8.5) aus Definition 8.3. Für $s^2 = 0$ ist $\int g(x,y) dP(x,y) = 0$ wegen der Stetigkeit von Q, also gilt (8.3) nicht. Sei im folgenden $s^2 > 0$. Jetzt folgt (8.3) aus $cs^2 > 0$. Mit $E_G(xx')$ ist auch

$$\int xx' g(x,y) dP(x,y) = \int 1(y^2 \leq cs^2) dQ(y) \int xx' dG(x)$$

invertierbar und $\int y^2 g(x,y) dP(x,y) \leq \int y^2 dQ(y) = 1$. Also sind für $dR = g dP$ die Voraussetzungen von Hilfssatz 11.1 erfüllt, wobei aus der G–Integrierbarkeit von $\|x\|^2$ auch die P– und R–Integrierbarkeit folgt. Damit gilt für g auch (8.4) und

$$\beta(g,P) = \arg\min_\beta \int \int (y - x'\beta)^2 g(x,y) dQ(y) dG(x) =$$
$$= \left[\int xx' g(x,y) dG(x) \right]^{-1} \int xyg(x,y) dQ(y) dG(x) = 0,$$
$$\text{da} \int xyg(x,y) dQ(y) dG(x) = \int y 1(y^2 < cs^2) dQ(y) \int x dG(x) = 0$$

wegen der Unabhängigkeit von x und y. Damit gilt

$$\sigma^2(g,P) = \frac{\int y^2 1(y^2 < cs^2) dQ(y)}{\int 1(y^2 < cs^2) dQ(y)},$$

d.h. $\sigma^2(g,P) = s^2 \Leftrightarrow j(s^2) = 0$. Für $\beta(g,P) = 0$ ist (8.5) genau dann erfüllt (und damit g KQ-FPCI bzgl. P), wenn $\sigma^2(g,P) = s^2$.

Beweis von Schritt 3: Angenommen, es gäbe einen KQ-FPCI g mit

$$g(x,y) = 1\left((y - x'\theta)^2 \le cs^2\right),\ s^2 \ge 0, \theta \ne 0.$$

Der Fall $s^2 = 0$ scheidet wieder aus, da dann (8.3) verletzt wäre. Nach (8.5) wäre

$$\theta = \arg\min_{\beta} \int\int (y - x'\beta)^2 1\left((y - x'\theta)^2 \le cs^2\right) dP(x,y). \tag{12.8}$$

Sei für $t \in I\!R^{p+1}$

$$F_\theta(t) := \int\int (y - x't)^2 1\left((y - x'\theta)^2 \le cs^2\right) dQ(y)dG(x). \tag{12.9}$$

Fall 1: $G(\{|x'\theta| \ge \sqrt{c}s + m\}) = 1$. Dann ist

$$\int\int 1\left((y - x'\theta)^2 \le cs^2\right) dQ(y)dG(x) = 0$$

wegen $\operatorname{supp} Q = [-m, m]$ im Widerspruch zu (8.3).

Fall 2: $G(\{x'\theta = 0\} | \{|x'\theta| < \sqrt{c}s + m\}) = 1$. Dann gilt

$$F_\theta(\theta) = \int\int y^2 1\left((y - x'\theta)^2 \le cs^2\right) dQ(y)dG(x) = F_\theta(0).$$

Also minimiert θ nicht eindeutig F_θ im Widerspruch zu (8.4).

Fall 3: $G(\{x'\theta \ne 0\} \cap \{|x'\theta| < \sqrt{c}s + m\}) > 0$. Mit $dR = g_{\theta,s^2} dP$, wobei $g_{\theta,s^2}(x,y) := 1((y - x'\theta)^2 \le cs^2)$, ist wieder Voraussetzung (11.1) von Hilfssatz 11.1 erfüllt. Also gilt mit (11.2):

$$v := \frac{\theta}{\|\theta\|}, \qquad \frac{\partial}{\partial v} F_\theta(t) := < v, \operatorname{grad} F_\theta(t) > =$$

$$= -2\sum_{i=1}^{p} v_i \int x_i \int (y - x't) 1\left((y - x'\theta)^2 \le cs^2\right) dQ(y)dG(x), \text{ also}$$

$$\frac{\partial}{\partial v} F_\theta(\theta) = -\frac{2}{\|\theta\|} \int x'\theta J(x'\theta) dG(x) \text{ mit}$$

$$J(u) := \int (y - u) 1\left((y - u)^2 \le c_0^2\right) dQ(y), \quad c_0 := \sqrt{cs^2}.$$

Ich zeige $uJ(u) < 0$ für alle u mit

$$0 < |u| < \sqrt{c}s + m. \tag{12.10}$$

Nach der Voraussetzung an G in Fall 3 folgt daraus $\int x'\theta J(x'\theta) dG(x) < 0$, also $\frac{\partial}{\partial v} F_\theta(\theta) > 0$ im Widerspruch zu (12.8), womit alles gezeigt wäre.

$$uJ(u) = \int u(y - u) 1(|y - u| \le c_0) q(y) dy =$$

$$= \int u|y - u| 1(|y - u| \le c_0)[1(y > u) - 1(y < u)] q(y) dy =$$

$$= \int u|t| 1(|t| \le c_0)[1(t > 0) - 1(t < 0)] q(t + u) dt =$$

$$= \int u|t| 1(0 < t \le c_0)[q(t + u) - q(-t + u)] dt =,$$

$$= \int |u| t 1(0 < t \le c_0)[q(t + |u|) - q(-t + |u|)] dt, \tag{12.11}$$

da für $u < 0$: $q(t + u) = q(-t + |u|)$, $q(-t + u) = q(t + |u|)$.

Nach Voraussetzung gilt $y \in (-m, m) \Rightarrow q(y) > 0$. Weiterhin ist q symmetrisch um 0 und streng monoton fallend zwischen 0 und m. Daher

$$t > 0, |u| > 0 \Rightarrow |u|t(q(t + |u|) - q(-t + |u|)) \leq 0.$$

Zusammen mit (12.11) ergibt sich

$$uJ(u) \leq \int |u|t1[\max(-m + |u|, 0) < \\ < t \leq \min(c_0, m + |u|)][q(t + |u|) - q(-t + |u|)]dt.$$

Letzteres ist kleiner als 0, denn mit (12.10) ist $-m + |u| < \min(c_0, m + |u|)$, weiter sind $c_0, m + |u| > 0$ und aus den Eigenschaften von q folgt für $t > 0$:

$$-m + |u| < t \leq m + |u| \Rightarrow q(-t + |u|) > 0 \Rightarrow q(t + |u|) - q(-t + |u|) < 0.$$

13 Fixpunktclusterindikatoren in Mischmodellen

In diesem Abschnitt werden Modelle der Form

$$(1-\epsilon)P(x,y)+\epsilon H^*(x,y) \tag{13.1}$$

behandelt, wobei P eine Verteilung ist, die zusammengehörige Punkte erzeugen soll, d.h. eine homogene Regressionsverteilung wie in den Sätzen 12.1 oder 12.4. Die genaue Form von H^* wird nicht festgelegt. H^* könnte als ausreißergenerierende Verteilung interpretiert werden oder auch eine Mischung weiterer homogener Regressionsverteilungen sein. Eine solche Verteilung würde dann Datensätze mit mehreren Clustern generieren. H^* muß allerdings von P „gut getrennt sein". In Abschnitt 13.1 bedeutet das, daß H^* im Bereich des Fixpunktclusters von P keine Masse hat, also alle Masse von H^* im Bereich der durch den FPC definierten Ausreißerregion liegt. Diese Resultate sind zum Beispiel auf die Modelle aus Abschnitt 2 nicht anwendbar, da dort der Störterm normalverteilt ist und daher auf ganz $\mathbb{R}$ eine nichtverschwindende Dichte hat.

In den darauffolgenden Abschnitten 13.2 und 13.3 werden überlappende Mischungen behandelt, wobei allerdings die Masse von H^* in den Bereichen, wo P „dicht" ist, stark beschränkt wird. Aufgrund rechnerischer Schwierigkeiten beschränke ich mich dort auf den Fall eindimensionaler Lokation (Abschnitt 13.2) und auf den Fall einer Regression ohne Achsenabschnitt (Abschnitt 13.3).

Alle Resultate sind Existenzresultate, d.h. die Eindeutigkeit der dort hergeleiteten FPCI wird nicht bewiesen. Eindeutigkeitsresultate wären aber auch anschaulich nicht sinnvoll, denn viele Mischmodelle der obigen Form erzeugen normalerweise mehrere anschauliche Cluster.

Keines der Resultate benötigt Voraussetzungen über Identifizierbarkeit. Das ist charakteristisch für die Fixpunktclusteranalyse (FPCA). Da die FPCA keine Optimallösung eines Entscheidungsproblems erzwingt, ist sie prinzipiell in der Lage, verschiedene Parametrisierungen desselben Modells zu finden (siehe dazu das Ende von Abschnitt 16.2 in den Simulationen). Allerdings beziehen sich die Beispiele für Identifizierbarkeitsprobleme in Abschnitt 5 auf Fälle der Mischungen von Regressionen mit normalverteiltem Störterm und Achsenabschnitt, die in den Sätzen dieses Abschnitts nicht behandelt werden. Die Spezialfälle, die in den Beispielen 13.6 und 13.14 diskutiert werden, haben identifizierbare Parameter: Auf eine Lokationsmischung von Normalverteilungen (Beispiel 13.6) kann Satz 6.7 mit $p=0$ angewendet werden. Mit demselben Satz folgt auch die Identifizierbarkeit einer Mischung von Regressionen ohne Achsenabschnitt mit normalverteilten Regressoren (Beispiel 13.14), denn der fehlende Achsenabschnitt bedeutet mit $\beta_{p+1}=0$ eine Einschränkung des Parameterraums. Dabei bleibt Identifizierbarkeit nach Bemerkung 4.7 erhalten.

13.1 Scharf trennbare Mischungen

In diesem Abschnitt wird Korollar 7.6 auf die Sätze über homogene Modelle angewendet, d.h. es werden Verteilungen H^* zu den Verteilungen aus den Sätzen 12.1 und 12.4 gemischt, die bezüglich der dort vorhandenen FPCI nur Ausreißer erzeugen. Satz 13.1 behandelt den Fall, daß zum homogenen Modell mit normalverteiltem Störterm eine Verteilung gemischt wird, die mit Wahrscheinlichkeit 1 Ausreißer gemäß Bemerkung 12.3

erzeugt. Satz 13.2 behandelt eine Mischung aus homogenen Modellen, deren Störterm beschränkten Träger wie in Satz 12.4 hat, so daß sich die von den verschiedenen homogenen Verteilungen erzeugten FPCI nicht überschneiden und damit jeweils füreinander Ausreißer sind.

In beiden Fällen existieren FPCI mit denselben Parametern wie in den Sätzen 12.1 und 12.4. Die Mischung mit H^* verzerrt also die Clusterindikatoren nicht.

Satz 13.1 (Normalverteilter Störterm und Ausreißer) *Sei* $c > 3$ *und*

$$R(x,y) = (1-\epsilon)P(x,y) + \epsilon H^*(x,y),$$

P definiert wie in Satz 12.1, $H^* \in \mathcal{P}_{p+2}$, $0 < \epsilon < 1$, *und mit k gemäß Satz 12.1 gelte*

$$H^*\{(x,y) : (y - x'\beta_0)^2 \leq ck\sigma_0^2\} = 0. \tag{13.2}$$

Dann existiert ein KQ-FPCI g mit

$$\beta(g,R) = \beta_0, \qquad \sigma^2(g,R) = k\sigma_0^2. \tag{13.3}$$

Beweis: Nach Korollar 7.6 ist $g(x,y) = 1[(y - x'\beta_0)^2 \leq ck\sigma_0^2]$ KQ-FPCI bzgl. R, wenn

$$g \text{ KQ-FPCI bzgl. } P \text{ ist,} \tag{13.4}$$

$$R(\{g=1\}) = (1-\epsilon)P(\{g=1\}) > 0, \tag{13.5}$$

$$R_{\{g=1\}} = P_{\{g=1\}}. \tag{13.6}$$

(13.4) folgt aus Satz 12.1. (13.5) gilt wegen $\epsilon < 1$. Weiter gilt $gdH^* \equiv 0$ wegen (13.2). Daher gilt (13.6) für $B \in I\!B^{p+2}$:

$$R_{\{g=1\}}(B) = \frac{\int_B g(x,y)d[(1-\epsilon)P+\epsilon H^*](x,y)}{\int g(x,y)d[(1-\epsilon)P+\epsilon H^*](x,y)} =$$

$$= \frac{\int_B g(x,y)dP(x,y)}{\int g(x,y)dP(x,y)} = P_{\{g=1\}}(B).$$

Nach Bemerkung 8.4 und der Definition von g gilt dann auch (13.3).

Satz 13.2 (Mischung mit beschränktem Störterm-Träger) *Es sei $c > 3$ und*

$$R(x,y) = \sum_{i=1}^{s} \epsilon_i P_i(x,y),$$

$$0 < \epsilon_i < 1, \sum_{i=1}^{s} \epsilon_i = 1,$$

$$P_i(x,y) = \int 1(t \le x) Q_i(y - t'\beta_i) dG_i(t), \qquad i = 1, \ldots, s,$$

wobei Q_i mit Riemann-Dichte q_i auf $[-m_i, m_i]$ sowie G_i für alle i die Voraussetzungen aus Satz 12.4 erfüllen. Außerdem gelte

$$\forall i \neq j : P_i\{(x,y) : (y - x'\beta_j)^2 \le c \operatorname{Var}(Q_j)\} = 0. \tag{13.7}$$

Dann existieren KQ-FPCI $g_i, i = 1, \ldots, s$ mit

$$\beta(g_i, R) = \beta_i, \qquad 0 < \sigma^2(g_i, R) \le \operatorname{Var}(Q_i) \ \forall i. \tag{13.8}$$

Beweis: Sei $i \in \{1, \ldots, s\}$ fest. Nach Satz 12.4 existiert ein KQ-FPCI

$$g_i = 1[(y - x'\beta_i)^2 \le c k_i \operatorname{Var}(Q_i)]$$

bzgl. P_i, k_i definiert wie k in Satz 12.4. Nach Korollar 7.6 ist g_i KQ-FPCI bzgl. R, wenn

$$R(\{g_i = 1\}) = \epsilon_i P_i(\{g_i = 1\}) > 0,$$
$$R_{\{g_i=1\}} = P_{i_{(g=1)}}.$$

Ersteres gilt wegen $\epsilon_i > 0$. Wegen (13.7) und $k_i < 1$ (Satz 12.4) ist $g_i dP_j \equiv 0 \ \forall j \neq i$, also auch

$$g_i dH^* \equiv 0 \text{ mit } dH^* := d\left[\sum_{j \neq i} \epsilon_j P_j\right].$$

Daher ist für $B \in \mathcal{B}^{p+2}$

$$R_{\{g_i=1\}}(B) = \frac{\int_B g_i(x,y) d[\epsilon_i P_i + H^*](x,y)}{\int g_i(x,y) d[\epsilon_i P_i + H^*](x,y)} =$$
$$= \frac{\int_B g_i(x,y) dP_i(x,y)}{\int g_i(x,y) dP_i(x,y)} = P_{i_{(g_i=1)}}(B).$$

Nach Bemerkung 8.4 und Definition von g gilt dann auch (13.8).

Abbildung 9 zeigt eine Situation, wie sie in Satz 13.2 behandelt wird. Dabei sind die dicken Linien die Regressionsgeraden der einzelnen Mischungskomponenten, die Kästen sind die Träger der einzelnen Verteilungen (auch die Regressoren haben in der Abbildung beschränkten Träger) und die gestrichelten Linien markieren die Ränder der Fixpunktcluster. Je nach Dichte des Störterms kann für gegebenen Regressor der Bereich des Fixpunktclusters zwischen $-ck_i \operatorname{Var}(Q_i)$ und $ck_i \operatorname{Var}(Q_i)$ breiter oder schmaler sein als der Träger der Störverteilung $[-m_i, m_i]$. Voraussetzung (13.7) besagt, daß der Bereich zwischen den gestrichelten Linien einer Verteilungskomponente nicht die Träger der anderen Verteilungskomponenten, d.h. die Kästen, schneiden darf.

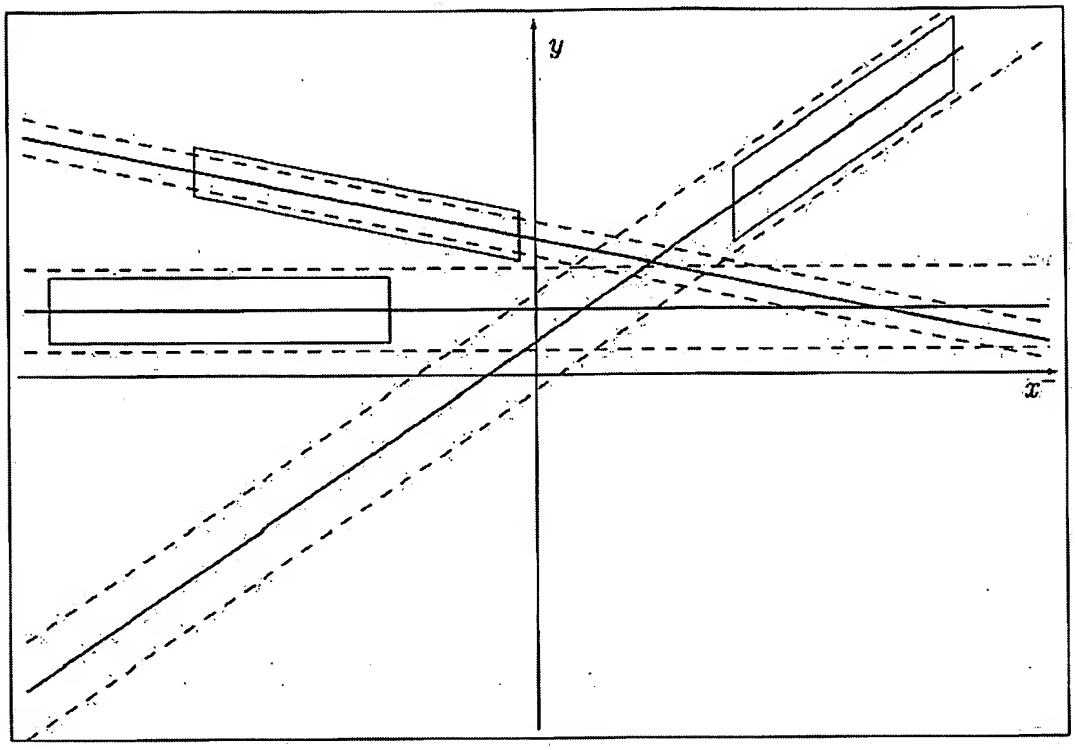


Abbildung 9: Beispiel für Satz 13.2

13.2 Überlappende Mischungen im Lokationsfall

In diesem Abschnitt wird ein erstes Resultat über die Existenz von Fixpunktclustern in einem Modell der Form (13.1) gezeigt, wobei sich die Verteilungen P und H^* überlappen. Ich behandle hier den eindimensionalen Lokationsfall, d.h. Q sei eine Mischung aus einer univariaten Normalverteilung P und irgendeiner anderen Verteilung H^* auf $\mathbb{R}$.

Bemerkung 13.3 (KQ-FPCI im Lokationsfall) *In der Schreibweise der Definition 8.3 und Bemerkung 8.4 bedeutet das: $p = 0, G(x) = \delta_1$. Sei $g_{u,s^2}(y) = 1((y-u)^2 \leq cs^2)$. Wegen $x \equiv 1, y^2 g_{u,s^2}(y) \leq (u + \sqrt{c}s)^2$ sind mit $dR = g_{u,s^2} dQ$ alle Voraussetzungen von Hilfssatz 11.1 erfüllt und*

$$\arg\min_{\beta} \int (y-\beta)^2 dR(y) = \frac{\int y dR(y)}{\int dR(y)}.$$

Also ist g_{u,s^2} genau dann KQ-FPCI bzgl. Q, wenn

$$\int g(y) dQ(y) > 0, \tag{13.9}$$

$$(u, s^2) \text{ ist Fixpunkt von } f = (\beta, \sigma^2) : \mathbb{R} \times \mathbb{R}_0^+ \mapsto \mathbb{R} \times \mathbb{R}_0^+,$$

$$\beta(u, s^2) = \frac{\int y g_{u,s^2}(y) dQ(y)}{\int g_{u,s^2}(y) dQ(y)},$$

$$\sigma^2(u, s^2) = \frac{\int (y - \beta(y,Q))^2 g_{u,s^2}(y) dQ(y)}{\int g_{u,s^2}(y) dQ(y)}. \tag{13.10}$$

Wenn $P = \mathcal{N}_{(\beta_0,\sigma_0^2)}$, dann besagt Satz 13.4, daß bzgl. Q gemäß (13.1) ein KQ-FPCI g existiert, so daß $\beta(g,Q) \in M_0$, $\sigma^2(g,Q) \in S_0$, wobei M_0 eine beschränkte Umgebung von β_0 und S_0 eine beschränkte Umgebung von σ_0^2 ist. Die wesentliche Voraussetzung des Satzes ist, daß

$$\forall I = \{y : (y-u)^2 \le cs^2,\ u \in M_0, s^2 \in S_0\} :\ H^*(I) \le \epsilon_0 P(I),$$

wobei ϵ_0 von ϵ abhängt. Diese Voraussetzung bedeutet, daß entweder H^* und P sehr gut voneinander getrennt sind - H^* hat wenig Masse, wo sich P „clustert" - oder daß ϵ sehr klein ist.

Die Parameter des FPCI entsprechen nur approximativ den Modellparametern, aber der Satz gibt konkrete Schranken an und es gilt $M_0 \to \{\beta_0\}$ für $\epsilon_0 \to 0, \epsilon \to 0$ oder $\sigma_0 \to 0$ (siehe Hilfssatz 13.7 und die Korollare 13.8 und 13.9).

In Beispiel 13.6 wird der Satz illustriert, indem in einer Mischung zweier Normalverteilungen die Mengen M_0 und S_0 berechnet werden.

Bezeichnungen und Konstanten für Satz 13.4: Es sei $P := \mathcal{N}_{(\beta_0,\sigma_0^2)}, \sigma_0^2 \ge 0$. Es werden eine Fixpunktcluster-Justierkonstante c sowie eine Konstante ϵ^* benötigt, so daß folgende Ungleichungen mit $q := 1.5$ erfüllt sind:

$$c_0 := 4.2974 = 2e^{\frac{1}{2}} + 1 < c < 25, \quad 0 < \epsilon^* < \frac{1}{(4c-1)q}. \tag{13.11}$$

Daraus folgt $(4c-1)q > 24.28$, also $\epsilon^* < 0.0412$, sowie

$$1 > c_1(\epsilon^*) := 1 - (4c-1)\epsilon^* > 1 - \frac{1}{q} = \frac{q-1}{q} > 0.$$

Es gelten die Bezeichnungen aus Abschnitt 11.2. Sei für $(u,s) \in \mathbb{R} \times \mathbb{R}^+$:

$$k(u,s,\epsilon^*) := E(u,s) - u + \epsilon^* s.$$

k ist offenbar stetig. Nach Hilfssatz 11.4 ist k streng monoton fallend in u und konvergiert mit $u \to \infty$ gegen $(\epsilon^*-1)s < 0$. Andererseits ist $k(0,s,\epsilon^*) = \epsilon^* s > 0$. Daher hat $k(\bullet,s,\epsilon^*)$ eine eindeutige Nullstelle $K_0(s,\epsilon^*) > 0$. Es seien

$$K_0^*(\epsilon^*) := \sup\{K_0(\sqrt{c}s,\epsilon^*) :\ \tfrac{1}{c} \le s^2 \le \tfrac{1}{c_1(\epsilon^*)}\},$$

(In Schritt 2 des Beweises zu Satz 13.4 wird $K_0^*(\epsilon^*) < \infty$ gezeigt.)

$$M_0(\epsilon^*) := [\beta_0 - \sigma_0 K_0^*(\epsilon^*), \beta_0 + \sigma_0 K_0^*(\epsilon^*)],$$
$$S_0(\epsilon^*) := \left[\tfrac{\sigma_0^2}{c}, \tfrac{\sigma_0^2}{c_1(\epsilon^*)}\right],$$
$$M_1(\epsilon^*) := \left[\inf M_0(\epsilon^*) - \sqrt{c \sup S_0(\epsilon^*)}, \sup M_0(\epsilon^*) + \sqrt{c \sup S_0(\epsilon^*)}\right] =$$
$$= \left[\beta_0 - \sigma_0\left(K_0^*(\epsilon^*) + \sqrt{\tfrac{c}{c_1(\epsilon^*)}}\right), \beta_0 + \sigma_0\left(K_0^*(\epsilon^*) + \sqrt{\tfrac{c}{c_1(\epsilon^*)}}\right)\right].$$

Damit kann nun der Satz formuliert werden:

Satz 13.4 (Lokationsmischung mit Normalverteilung) *Es sei mit* $0 < \epsilon < 1$

$$Q(y) := (1-\epsilon)P(y) + \epsilon H^*(y),$$

$H^* \in \mathcal{P}_1$ *sei stetig auf* $M_1(\epsilon^*)$. *Für* $\epsilon_0 := \frac{\epsilon^*(1-\epsilon)}{\epsilon}$ *gelte*

$$H^*[a,b] \leq \epsilon_0 P[a,b] \tag{13.12}$$

$$\forall [a,b] = [m - \sqrt{c}s, m + \sqrt{c}s] \text{ mit } m \in M_0(\epsilon^*), s^2 \in S_0(\epsilon^*). \tag{13.13}$$

Dann existiert ein KQ-FPCI g *bzgl.* Q *mit*

$$\beta(g,Q) \in M_0(\epsilon^*),\ \sigma^2(g,Q) \in S_0(\epsilon^*). \tag{13.14}$$

Beweis: Im Falle $\sigma_0^2 = 0$ ist $\int 1(y^2 = 0)dQ(y) \geq (1-\epsilon) > 0$ und

$$\int (y-u)^2 1(y^2 = 0)dQ(y) = 0 \Leftrightarrow u = 0,$$

so daß $\beta(0,0) = 0 = \sigma^2(0,0)$. Damit sind für $g_{0,0}$ (13.9) und (13.10) gezeigt, es ist KQ-FPCI bzgl. Q, und der Satz folgt mit $M_0(\epsilon^*) = S_0(\epsilon^*) = \{0\}$.

Sei nun $\sigma_0^2 > 0$. Aufgrund der Äquivarianzeigenschaften der FPCI (Bemerkung 8.5) sei ohne Beschränkung der Allgemeinheit $\beta_0 = 0, \sigma_0^2 = 1$. Sei f definiert wie in Bemerkung 13.3. Sei $M := M_0(\epsilon^*) \times S_0(\epsilon^*)$. Angenommen, $(u,s^2) \in M$ sei Fixpunkt von f, dann ist g_{u,s^2} mit Schritt 1 der folgenden Argumentation auch KQ-FPCI für Q. Weiterhin wird gezeigt: Die Einschränkung von f auf M ist eine Selbstabbildung (Schritt 2-5 zeigen: $(u,s^2) \in M \Rightarrow f(u,s^2) \in M$) und stetig (Schritt 6). Brouwers Fixpunktsatz (zum Beispiel Satz 229.2 aus Heuser (1981)) sichert die Existenz eines Fixpunktes einer stetigen Selbstabbildung einer kompakten und konvexen nichtleeren Teilmenge des $\mathbb{R}^d$. $M_0(\epsilon^*)$ und $S_0(\epsilon^*)$ sind abgeschlossene Intervalle, wobei $S_0(\epsilon^*)$ nach Definition und $M_0(\epsilon^*)$ nach Schritt 2 kompakt sind. Also ist auch M ist als Produkt kompakter Intervalle kompakt und konvex und wegen $K_0^*(\epsilon^*) > 0 \Rightarrow (\beta_0, \sigma_0^2) \in M$ nichtleer und der Beweis liefert die Existenz eines Fixpunktes von f auf M. Folgende Behauptungen werden gezeigt:

Schritt 1: $(u,s^2) \in M \Rightarrow g_{u,s^2}$ erfüllt (13.9).

Schritt 2:

$$M_0(\epsilon^*) \subset [-0.6252, 0.6252]$$

Schritt 3:

$$\forall (u,s^2) \in M : \beta(u,s^2) \in M_0(\epsilon^*)$$

Schritt 4:

$$\forall (u,s^2) \in M : \sigma^2(u,s^2) \leq \frac{1}{c_1(\epsilon^*)}$$

Schritt 5:

$$\forall (u, s^2) \in M : \sigma^2(u, s^2) \geq \frac{1}{c}$$

Schritt 6: f ist stetig auf M.

Beweis von Schritt 1: $(u, s^2) \in M \Rightarrow s^2 > 0$. Damit:

$$\int g_{u,s^2}(y) dQ(y) \geq (1-\epsilon) \int g_{u,s^2}(y) d\mathcal{N}_{(\beta_0, \sigma_0^2)}(y) > 0.$$

Beweis von Schritt 2: Es gelten die oben eingeführten Bezeichnungen und Konstanten. Betrachte

$$\sup M_0(\epsilon^*) = \sup_{s^2 \in S_0(\epsilon^*)} K_0(\sqrt{c}s, \epsilon^*) = \sup_{s^2 \in S_0(\epsilon^*)} \{u : k(u, \sqrt{c}s, \epsilon^*) = 0\}.$$

Nach Definition hat k nur positive Nullstellen. $M_0(\epsilon^*)$ ist nach Definition symmetrisch um 0. Daher wird im folgenden immer $u > 0$ vorausgesetzt. Es gilt

$$c_1(\epsilon^*) > c_1\left(\frac{1}{(4c-1)q}\right) = \frac{q-1}{q},$$

so daß

$$S_0(\epsilon^*) \subset S_0\left(\tfrac{1}{(4c-1)q}\right) = \left[\tfrac{1}{c}, \tfrac{q}{q-1}\right] \subset \left[\tfrac{1}{25}, \tfrac{q}{q-1}\right] =: S_0^*,$$
$$\text{also } s^2 \in S_0^* \Leftrightarrow \sqrt{c}s \in C_0^* := \left[1, \sqrt{\tfrac{25q}{q-1}}\right].$$

Weiterhin ist k streng monoton steigend in ϵ^* und streng monoton fallend in u, und daher gilt mit $\epsilon^* < \frac{1}{(4c-1)q}$:

$$k(u, s, \epsilon^*) = 0 \Rightarrow k\left(u, s, \frac{1}{(4c-1)q}\right) > 0, \text{ also } u = K_0(s, \epsilon^*) < K_0\left(s, \frac{1}{(4c-1)q}\right).$$

Sei jetzt

$$k_0(u, s, c) := k\left(u, s, \frac{1}{(4c-1)q}\right) = E(u, s) + \frac{s}{(4c-1)q} - u.$$

k_0 fällt streng monoton in $c > c_0 = 4.2974$ und wegen Hilfssatz 11.4 in u. Also

$$k_0(u, s, c) = 0 \Rightarrow k_0(u, s, c_0) > 0, \text{ daher } u < K_0\left(s, \frac{1}{(4c_0-1)q}\right).$$

Zusammengesetzt:

$$\sup M_0(\epsilon^*) \leq \sup_{s \in C_0^*} \{u : k_0(u, s, c) = 0\} \leq \sup_{s \in C_0^*} \{u : k_0(u, s, c_0) = 0\}.$$

Es sei $u(s)$ definiert gemäß $k_0(u(s), s, c_0) \stackrel{!}{=} 0$. Nach Definition von k ist $u(s) > 0$ damit wohldefiniert. $u(s)$ soll nun für $s \in C_0^*$ maximiert werden.

k_0 ist stetig differenzierbar nach u und s. Wie sich aus Hilfssatz 11.4 ergibt, ist für beliebiges $s > 0$

$$\left[\frac{\partial k_0(u,s,c_0)}{\partial u}\bigg|_{u=u(s)}\right] < 0.$$

Damit kann der Satz über Differenzierung impliziter Funktionen (zum Beispiel Satz 170.1 aus Heuser (1981)) angewendet werden:

$$u'(s) = -\left[\frac{\partial k_0(u,s,c_0)}{\partial u}\bigg|_{u=u(s)}\right]^{-1}\left[\frac{\partial k_0(u,s,c_0)}{\partial s}\bigg|_{u=u(s)}\right] =$$
$$= -\frac{\left[u(s)E_+(u(s),s)-[s+E_+(u(s),s)]E(u(s),s)+\frac{1}{(4c_0-1)q}\right]}{\frac{\partial}{\partial u}E(u,s)\big|_{u=u(s)}{}^{-1}}$$

mit Hilfe von Hilfssatz 11.3. Alle Terme des Zählers sind (bis auf die angegebenen Vorzeichen) positiv. Da $s \in C_0^*$, gilt $s \geq 1$, also $E_+(u(s),s) < u(s)+1$ mit Hilfssatz 11.5. Damit läßt sich sowohl die Summe der Terme mit „+" als auch die Summe der Terme mit „−" in der großen Klammer durch

$$(s + E_+(u(s),s))u(s) + \frac{1}{(4c_0-1)q} < (s+u(s)+1)u(s) + \frac{1}{(4c_0-1)q}$$

abschätzen. Sei vorerst $u(s) < 0.63$ vorausgesetzt. Damit

$$\sup C_0^* = 15 \Rightarrow |\text{Zähler}| < 0.63(15+1.63) + 0.0412.$$

Der Nenner ist nach Hilfssatz 11.7 gleich $-V(u(s),s)$, was nach Hilfssatz 11.8 für $s \in C_0^*$ keinen kleineren Betrag hat als $V(u(s),1)$. Um den Nenner zu minimieren, betrachte man mit Hilfssatz 11.3

$$\frac{\partial}{\partial u}V(u,1) =$$

$$= 2uE_+(u,1) + 3uE^2(u,1) - 3E(u,1)E_+(u,1) - E(u,1)u^2 - 2E^3(u,1) \quad (13.15)$$

$$\Rightarrow \left|\frac{\partial}{\partial u}V(u,1)\right| < 9, \quad (13.16)$$

denn alle Terme in (13.15) sind (bis auf das angegebene Vorzeichen) positiv. Ich habe u hier immer durch 1 und $E_+(u,1)$ gemäß Hilfssatz 11.5 durch 2 abgeschätzt. Damit ist die Summe der Terme mit „+" kleiner als 7, die Summe der Terme mit „−" kleiner als 9. Daher überschätzt das Minimum von $V(u,1)$ aus 71 äquidistanten Stützstellen im Abstand 0.009 zwischen 0 und 0.63 um höchstens $9 * \frac{0.009}{2} = 0.0405$ das Minimum von $|-V(u,1)|$ für $|u| < 0.63$. Zusammen ergibt sich

$$|\text{Nenner}| > \min_{u<0.63}|V(u,1)| > 0.2337 \Rightarrow |u'(s)| < 45.01.$$

Damit kann nun $\max u(s)$ für $\sqrt{c}s \in C_0^* = [1,15]$ abgeschätzt werden: Das Maximum M_u von $u(s)$ an 41967 äquidistanten Stützstellen mit Abstand 0.000334 unterschätzt $\max u(s)$ um höchstens $45.01 * \frac{0.000334}{2} = 0.0075$. Mit $M_u = 0.6177$ ergibt sich $u(s) < 0.6252$.

Vorausgesetzt war dafür $u(s) < 0.63$. Diese Voraussetzung ist wegen $M_u = 0.6177 < 0.63$ für alle Stützstellen $1 = s_1, \ldots, s_{41967} = 15$ erfüllt. Angenommen, es gäbe nun $z \in [1, 15]$ mit $u(z) \geq 0.63$. Dann

$$\exists i \in 1, \ldots, 41967 : \; z \in \left(s_i - \frac{0.000334}{2}, s_i + \frac{0.000334}{2}\right].$$

Sei $z_0 := \inf\{z \in (s_i - \frac{0.000334}{2}, s_i + \frac{0.000334}{2}] : u(z) \geq 0.63\}$. Wegen der Stetigkeit von u ist $u(z_0) = 0.63$. Nach dem Mittelwertsatz der Differentialrechnung muß es dann $z_1 \in (s_i, z_0)$ geben mit

$$|u'(z_1)| = \left|\frac{u(z_0) - u(s_i)}{z_0 - s_i}\right| \geq \frac{0.63 - 0.6177}{0.000167} = 73.65.$$

Für s mit $u(s) < 0.63$ war aber $|u'(s)| < 45.01$, also $u(z_1) \geq 0.63$ im Widerspruch zur Definition von z_0. Also gilt $u(s) < 0.63$ für $s \in [1, 15]$ und damit auch $u(s) < 0.6252$.

Beweis von Schritt 3: Definiere

$$Q_- := (1-\epsilon)P + \epsilon H_-, \; H_- := \mathcal{L}(-y) \text{ für } \mathcal{L}(y) = H^*.$$

H_- erfüllt die Voraussetzungen des Satzes an H^* ebenfalls, da $M_0(\epsilon^*)$ und $M_1(\epsilon^*)$ symmetrisch um 0 sind und für alle $a > b \in \mathbb{R}$: $H_-[-b, -a] = H^*[a, b]$. Dann

$$\beta(g_{u,s^2}, Q) = \frac{(1-\epsilon)\int y g_{u,s^2}(y)dP(y) + \epsilon \int y g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y) + \epsilon \int g_{u,s^2}(y)dH^*(y)} =$$

$$= -\frac{(1-\epsilon)\int y g_{-u,s^2}(y)dP(y) + \epsilon \int y g_{-u,s^2}(y)dH_-(y)}{(1-\epsilon)\int g_{-u,s^2}(y)dP(y) + \epsilon \int g_{-u,s^2}(y)dH_-(y)} = -\beta(g_{-u,s^2}, Q_-),$$

$$\sigma^2(g_{u,s^2}, Q) = \sigma^2(g_{-u,s^2}, Q_-)$$

mit analoger Rechnung, denn P und die Menge $M_0(\epsilon^*)$ sind symmetrisch um 0. Sei daher ohne Einschränkung $u \geq 0$. Es gilt:

$$|\beta(u, s^2)| = \left|\frac{(1-\epsilon)\int y g_{u,s^2}(y)dP(y) + \epsilon \int y g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y) + \epsilon \int g_{u,s^2}(y)dH^*(y)}\right| \leq$$

(Der Nenner ist > 0 nach Schritt 1, die erste Hälfte des Zählers ist ≥ 0 wegen Hilfssatz 11.5.)

$$\leq \frac{(1-\epsilon)\int y g_{u,s^2}(y)dP(y) + \epsilon\left|\int y g_{u,s^2}(y)dH^*(y)\right|}{(1-\epsilon)\int g_{u,s^2}(y)dP(y) + \epsilon \int g_{u,s^2}(y)dH^*(y)} \leq \tag{13.17}$$

(Für $g_{u,s^2}(y) = 1 \Leftrightarrow y \in [u - \sqrt{c}s, u + \sqrt{c}s]$ gilt $y \leq u + \sqrt{c}s$.)

$$\leq \frac{(1-\epsilon)\int y g_{u,s^2}(y)dP(y) + \epsilon(u+\sqrt{c}s)\int g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y) + \epsilon \int g_{u,s^2}(y)dH^*(y)} \leq$$

(Diese Ungleichung wird am Ende von Schritt 3 gezeigt.)

$$\leq \frac{(1-\epsilon)\int y g_{u,s^2}(y)dP(y) + \epsilon c_0(u+\sqrt{c}s)\int g_{u,s^2}(y)dP(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y) + \epsilon c_0 \int g_{u,s^2}(y)dP(y)} = \tag{13.18}$$

(Herauskürzen von $(1-\epsilon)\int g_{u,s^2}(y)dP(y)$, einsetzen von $\epsilon^* = \frac{\epsilon\epsilon_0}{1-\epsilon}$.)

$$= \frac{E(u,\sqrt{cs})+\epsilon^*(u+\sqrt{cs})}{1+\epsilon^*}.$$

Das heißt:

$$\frac{E(u,\sqrt{cs})+\epsilon^*(u+\sqrt{cs})}{1+\epsilon^*} - u \le 0 \Rightarrow |\beta(u,s^2)| - u \le 0.$$

Durch Multiplikation der linken Seite mit $(1+\epsilon^*)$ ist das gleichbedeutend mit der Inklusion

$$k(u,\sqrt{cs},\epsilon^*) = E(u,\sqrt{cs}) + \epsilon^*\sqrt{cs} - u \le 0 \Rightarrow |\beta(u,s^2)| \le u. \tag{13.19}$$

$K_0(\sqrt{cs},\epsilon^*) > 0$ ist die eindeutige Nullstelle von $k(\bullet,\sqrt{cs},\epsilon^*)$. Sei

$$v := K_0^*(\epsilon^*) = \max M_0(\epsilon^*) \le 0.6252.$$

Schritt 3 ist bewiesen, wenn

$$u \le v, s^2 \in S_0(\epsilon^*) \Rightarrow |\beta(u,s^2)| \le v.$$

Da k streng monoton in u fällt, gilt

$$s^2 \in S_0(\epsilon^*) \Rightarrow k(v,\sqrt{cs},\epsilon^*) \le k\,[K_0(\sqrt{cs},\epsilon^*),\sqrt{cs},\epsilon^*] = 0 \Rightarrow$$
$$\Rightarrow \frac{E(v,\sqrt{cs})+\epsilon^*\sqrt{cs}-v}{1+\epsilon^*} \le 0 \Rightarrow \frac{E(v,\sqrt{cs})+\epsilon^*(v+\sqrt{cs})}{1+\epsilon^*} \le v.$$

Aus Hilfssatz 11.7 geht hervor, daß $E(u,s)$ monoton in u steigt. Damit und mit der Ungleichungskette zu Beginn des Beweises von Schritt 3 folgt:

$$\forall u \le v: \ |\beta(u,s^2)| \le \frac{E(u,\sqrt{cs})+\epsilon^*(u+\sqrt{cs})}{1+\epsilon^*} \le$$
$$\le \frac{E(v,\sqrt{cs})+\epsilon^*(v+\sqrt{cs})}{1+\epsilon^*} \le v.$$

Es bleibt die Ungleichung zu zeigen, die zu (13.18) führt. Zuerst wird ein Hilfsresultat bewiesen:

$$d > 0, a,b,e,h \ge 0, h \ge \frac{a}{d}, e \ge b \Rightarrow \frac{a+bh}{d+b} \le \frac{a+eh}{d+e}. \tag{13.20}$$

Beweis von (13.20): Aus den Voraussetzungen folgt $a(e-b) \le hd(e-b)$. Damit ist $ae+bhd \le ab+ehd$. Weiter gilt

$$\frac{a+bh}{d+b} = \frac{ad+bhe+ae+bhd}{(d+b)(d+e)} \le \frac{ad+bhe+ab+ehd}{(d+b)(d+e)} = \frac{a+eh}{d+e}.$$

Es seien nun

$$a := (1-\epsilon)\int y g_{u,s^2}(y)dP(y), \quad b := \epsilon\int g_{u,s^2}(y)dH^*(y),$$
$$d := (1-\epsilon)\int g_{u,s^2}(y)dP(y), \quad e := \epsilon\epsilon_0\int g_{u,s^2}(y)dP(y),$$
$$h := (u+\sqrt{cs}).$$

$a \geq 0$ gilt wegen $u \geq 0 \Rightarrow E(u,s) \geq 0$ (Hilfssatz 11.5). Voraussetzung (13.12) garantiert $e \geq b$, denn $\{y : g_{u,s^2}(y) = 1\}$ ist ein Intervall der Form (13.13). Hilfssatz 11.5 erbringt

$$0 \leq \frac{a}{d} = E(u, \sqrt{c}s) \leq u + \sqrt{c}s = h,$$

Schritt 1 bringt $d > 0$, also sind die Voraussetzungen von (13.20) erfüllt. Daraus ergibt sich die Ungleichung (13.18).

Beweis von Schritt 4:

$$\sigma^2(u,s^2) = \frac{(1-\epsilon)\int[y-\beta(u,s^2)]^2 g_{u,s^2}(y)dP(y)+\epsilon\int[y-\beta(u,s^2)]^2 g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y)+\epsilon\int g_{u,s^2}(y)dH^*(y)} \leq \quad (13.21)$$

($\beta(u,s^2)$ minimiert (13.21) nach Hilfssatz 11.1 mit $x \equiv 1$, $dR = g_{u,s^2}dQ$.)

$$\leq \frac{(1-\epsilon)\int[y-E(u,\sqrt{c}s)]^2 g_{u,s^2}(y)dP(y)+\epsilon\int[y-E(u,\sqrt{c}s)]^2 g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y)+\epsilon\int g_{u,s^2}(y)dH^*(y)} \leq$$

(Aus $g_{u,s^2}(y) = 1 \Leftrightarrow y \in [u - \sqrt{c}s, u + \sqrt{c}s]$ folgt $[y - E(u,\sqrt{c}s)]^2 \leq 4cs^2$, da auch $g_{u,s^2}(E(u,\sqrt{c}s)) = 1$ nach Hilfssatz 11.5.)

$$\leq \frac{(1-\epsilon)\int[y-E(u,\sqrt{c}s)]^2 g_{u,s^2}(y)dP(y)+\epsilon 4cs^2\int g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y)+\epsilon\int g_{u,s^2}(y)dH^*(y)} =: B.$$

Nun soll (13.20) angewendet werden. Es seien

$$a := (1-\epsilon)\int[y - E(u,\sqrt{c}s)]^2 g_{u,s^2}(y)dP(y), \quad b := \epsilon\int g_{u,s^2}(y)dH^*(y),$$
$$d := (1-\epsilon)\int g_{u,s^2}(y)dP(y), \quad e := \epsilon\epsilon_0\int g_{u,s^2}(y)dP(y),$$
$$h := 4cs^2.$$

Es sind offenbar wieder $a, b, e, h \geq 0, d > 0, e \geq b$ und nach Definition von $V(u,s)$:

$$0 \leq \frac{a}{d} = V(u,\sqrt{c}s) \leq 4cs^2 = h,$$

also sind die Voraussetzungen von (13.20) erfüllt. Daraus ergibt sich

$$B \leq \frac{(1-\epsilon)\int(y-E(u,\sqrt{c}s))^2 g_{u,s^2}(y)dP(y)+\epsilon\epsilon_0 4cs^2\int g_{u,s^2}(y)dP(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y)+\epsilon\epsilon_0\int g_{u,s^2}(y)dP(y)} =$$

(Herauskürzen von $(1-\epsilon)\int g_{u,s^2}(y)dP(y)$ und einsetzen von $\epsilon^* = \frac{\epsilon\epsilon_0}{1-\epsilon}$; die letzte Ungleichung gilt nach Hilfssatz 11.7.)

$$= \frac{V(u,\sqrt{c}s)+\epsilon^* 4cs^2}{1+\epsilon^*} < \frac{1+\epsilon^* 4cs^2}{1+\epsilon^*} =: l(s^2).$$

Sei nun

$$m(s^2) := l(s^2) - s^2 = \left(\frac{4\epsilon^* c}{1+\epsilon^*} - 1\right)s^2 + \frac{1}{1+\epsilon^*}.$$

Für $s_0^2 := \frac{1}{c_1(\epsilon^*)} = \frac{1}{1+(1-4c)\epsilon^*}$ ist

$$m(s_0^2) = \frac{4\epsilon^* c - (1+\epsilon^*) + 1 + (1-4c)\epsilon^*}{(1+\epsilon^*)(1+(1-4c)\epsilon^*)} = 0.$$

Die Aussage von Schritt 4 ist zu zeigen, also

$$\sigma^2(u, s^2) \leq s_0^2 \qquad \forall s^2 \leq s_0^2. \tag{13.22}$$

Beweis von (13.22): Mit Voraussetzung (13.11) ergibt sich

$$\frac{1}{4\epsilon^*} + \frac{1}{4} > \frac{1}{4q\epsilon^*} + \frac{1}{4} > c \Rightarrow \frac{1+\epsilon^*}{4c\epsilon^*} > 1 \Rightarrow \frac{4c\epsilon^*}{1+\epsilon^*} - 1 < 0.$$

Daher ist m streng monoton fallend in s^2. Weiterhin ist $m(0) = \frac{1}{1+\epsilon^*} > 0$, also ist $m(s^2) \geq 0 \quad \forall s^2 \leq s_0^2$. Schließlich steigt l streng monoton in s^2. Daraus folgt (13.22) wie folgt: (Die erste Ungleichung ist die Ungleichungskette vom Anfang des Beweises von Schritt 4, die letzte Gleichung ist äquivalent zu $m(s_0^2) = 0$.)

$$\sigma^2(u, s^2) \leq l(s^2) \leq l(s_0^2) = s_0^2 \qquad \forall s^2 \leq s_0^2.$$

Beweis von Schritt 5: Es sei wieder ohne Einschränkung $u \geq 0$, also $u \geq E(u, s) \geq 0$ mit Hilfssatz 11.5. Der Beweis hat zwei Teilschritte:

Schritt 5a:

$$\sigma^2(u, s^2) \geq \frac{1}{1+\epsilon^*} V(u, \sqrt{c}s)$$

Schritt 5b:

$$\frac{1}{1+\epsilon^*} V(u, 1) \geq \frac{1}{c}.$$

Weil $V(u, s)$ nach Hilfssatz 11.8 für $u < 0.63$ und $s \geq 1$ (also nach Schritt 2 für alle $(u, \sqrt{c}s)$ mit $(u, s^2) \in M$) in s streng monoton steigt, ergibt sich zusammengesetzt:

$$s^2 \geq \frac{1}{c} \Rightarrow \sigma^2(u, s^2) \geq \frac{1}{1+\epsilon^*} V(u, 1) \geq \frac{1}{c}.$$

Beweis von Schritt 5a:

$$\sigma^2(u, s^2) = \frac{(1-\epsilon)\int (y-\beta(u,s^2))^2 g_{u,s^2}(y)dP(y) + \epsilon \int (y-\beta(u,s^2))^2 g_{u,s^2}(y)dH^*(y)}{(1-\epsilon)\int g_{u,s^2}(y)dP(y) + \epsilon \int g_{u,s^2}(y)dH^*(y)} \geq$$

(Der zweite Summand des Zählers wird durch 0 abgeschätzt, der zweite Summand des Nenners nach Voraussetzung (13.12) durch $\epsilon\epsilon_0 \int g_{u,s^2}(y)dP(y)$.)

$$\geq \frac{(1-\epsilon)\int (y-\beta(u,s^2))^2 g_{u,s^2}(y)dP(y)}{(1-\epsilon+\epsilon\epsilon_0)\int g_{u,s^2}(y)dP(y)} =$$

($\epsilon^* = \frac{\epsilon\epsilon_0}{1-\epsilon}$; zuletzt Ersetzung von $\beta(u, s^2)$ durch $E(u, \sqrt{c}s)$, was den Ausdruck nach Hilfssatz 11.1 mit $x \equiv 1$, $dR = g_{u,s^2}dP$ minimiert.)

$$= \frac{(1-\epsilon)\int (y-\beta(u,s^2))^2 g_{u,s^2}(y)dP(y)}{(1-\epsilon)(1+\epsilon^*)\int g_{u,s^2}(y)dP(y)} \geq \frac{1}{1+\epsilon^*} V(u, \sqrt{c}s).$$

Beweis von Schritt 5b: Mit $c > c_0$, $0 < \epsilon^* < \frac{1}{(4c-1)q} =: m$, $\sqrt{c}s \geq 1$ und Hilfssatz 11.8 ist

$$\frac{V(u,\sqrt{c}s)}{1+\epsilon^*} - \frac{1}{c} \geq h(u) := \frac{V(u,1)}{1+m} - \frac{1}{c_0}. \tag{13.23}$$

Unter Verwendung von (13.16) folgt

$$|h'(u)| < \left|\frac{\partial}{\partial u}V(u,1)\right| < 9.$$

Das heißt, daß $\min_{u\leq 0.6252} h(u)$ durch das Minimum von $h(u)$ an 626 äquidistanten Stützstellen mit Abstand 0.001 um höchstens $9 * \frac{0.001}{2} = 0.0045$ überschätzt wird. Damit ergibt sich

$$\min_{u\leq 0.6252} h(u) \geq 0.0263 > 0.$$

Mit (13.23) ergibt sich Schritt 5b.

Beweis von Schritt 6: Schritt 6 folgt aus Hilfssatz 11.2 mit $\theta = u$. In diesem Fall ist $x \equiv 1$, womit die Voraussetzung (11.5) direkt und (11.8) mit Schritt 1 folgt. Voraussetzung (11.7) ist genau Schritt 1 und (11.6) folgt aus der Stetigkeit von H^* auf $M_1(\epsilon^*)$.

Bemerkung 13.5 *Satz 13.4 könnte auch mit einem anderen $q > 1$ und anderen Schranken für c formuliert werden. Allerdings würde eine Vergrößerung des Bereiches der zulässigen c zur Folge haben, daß q größer und damit nach (13.11) ϵ^* kleiner gewählt werden müßte, so daß die Voraussetzung (13.12) des Satzes schärfer würde und der Satz auf weniger Verteilungen angewandt werden könnte.*

Beispiel 13.6 *Es gelten die Bezeichnungen von Satz 13.4. Es sei*

$$Q := \frac{1}{2}\mathcal{N}_{(a_1,1)} + \frac{1}{2}\mathcal{N}_{(a_2,1)}.$$

Sei $d := |a_1 - a_2|$. Für $c = 10$, $d \geq 6.108$ existieren KQ-FPCI g_i, $i = 1, 2$, bzgl. Q mit

$$|\beta(g_i, Q) - a_i| \leq 0.0139, \qquad 0.1 \leq \sigma^2(g_i, Q) \leq 1.185.$$

Für $c = 1.02(2e^{\frac{1}{2}} + 1) = 4.383$ und $d \geq 4.642$ existieren KQ-FPCI $g_i, i = 1, 2$, bzgl. Q mit

$$|\beta(g_i, Q) - a_i| \leq 0.0515, \qquad 0.228 \leq \sigma^2(g_i, Q) \leq 1.330.$$

Beweis: Es ist zu zeigen, daß die Voraussetzungen des Satzes 13.4 erfüllt sind. Offenbar gilt (13.11) für $c = 10$ und $c = 4.383$. Es folgt $\epsilon^* < 0.017$ bzw. $\epsilon^* < 0.040$. Sei ohne Einschränkung

$$P = \mathcal{N}(a_1, 1),\ H^* = \mathcal{N}(a_2, 1),\ a_1 = 0,\ a_2 = d > 0.$$

H^* ist stetig. Also ist nur noch Voraussetzung (13.12) nachzuprüfen. Es sei

$$J(m, s, d) := \frac{\Phi(m+s-d) - \Phi(m-s-d)}{\Phi(m+s) - \Phi(m-s)}.$$

Damit ist $\frac{H^*[m-s,m+s]}{P[m-s,m+s]} = J(m,s,d)$. (13.12) ist also erfüllt, wenn

$$\epsilon_0 \geq \max_{m\in M_0(\epsilon^*),s^2\in S_0(\epsilon^*)} J(m,\sqrt{c}s,d). \tag{13.24}$$

Wegen $\epsilon = \frac{1}{2}$ ist $\epsilon_0 = \epsilon^*$. Es gelten

$$\frac{\partial}{\partial m}J(m,s,d) = \frac{[\varphi(m+s-d)-\varphi(m-s-d)][\Phi(m+s)-\Phi(m-s)]-[\varphi(m+s)-\varphi(m-s)][\Phi(m+s-d)-\Phi(m-s-d)]}{[\Phi(m+s)-\Phi(m-s)]^2},$$
$$\frac{\partial}{\partial s}J(m,s,d) = \frac{[\varphi(m+s-d)+\varphi(m-s-d)][\Phi(m+s)-\Phi(m-s)]-[\varphi(m+s)+\varphi(m-s)][\Phi(m+s-d)-\Phi(m-s-d)]}{[\Phi(m+s)-\Phi(m-s)]^2},$$
$$\frac{\partial}{\partial d}J(m,s,d) = \frac{-\varphi(m+s-d)+\varphi(m-s-d)}{\Phi(m+s)-\Phi(m-s)}.$$

Durch Multiplikation mit $\Phi(m+s)-\Phi(m-s)$ und Division durch $\Phi(m+s-d)-\Phi(m-s-d)$ folgt

$$\frac{\partial}{\partial m}J(m,s,d) > 0 \Leftrightarrow E(m,s) - E(m-d,s) > 0.$$

Letzteres gilt wegen $d > 0$ und Hilfssatz 11.7. Also wird $J(m,s,d)$ maximiert, indem m maximal gewählt wird, also insbesondere $m > 0$. Entsprechend

$$\frac{\partial}{\partial s}J(m,s,d) > 0 \Leftrightarrow E_+(m-d,s) - E_+(m,s) > 0.$$

Nach Schritt 2 des Beweises zu Satz 13.4 gilt $m \in M_0(\epsilon^*) \Rightarrow m < 0.63$, also $d - m > \max M_0(\epsilon) > 0$. Also mit (11.11) und Hilfssatz 11.6: $E_+(m-d,s) = E_+(d-m,s) > E_+(m,s)$. Daher wird $J(m,s,d)$ maximiert, indem s maximal gewählt wird. Zuletzt ist

$$\frac{\partial}{\partial d}J(m,s,d) < 0 \Leftrightarrow E(m-d,s) < 0.$$

Das ist der Fall wegen Hilfssatz 11.5 und $m - d < 0$. Das bedeutet:

$$\epsilon_0 \geq \max_{m\in M_0(\epsilon^*),s^2\in S_0(\epsilon^*)} J(m,\sqrt{c}s,d) \Rightarrow \forall d_0 \geq d > 0 : \epsilon_0 \geq \max_{m\in M_0(\epsilon^*),s^2\in S_0(\epsilon^*)} J(m,\sqrt{c}s,d_0).$$

Für gegebenes ϵ^* sind also $M_0(\epsilon^*)$, $S_0(\epsilon^*)$ zu ermitteln. Dann sind die Voraussetzungen von Satz 13.4 erfüllt für $d_0 \geq d$ mit

$$\epsilon_0 = \epsilon^* \stackrel{!}{=} J[\max M_0(\epsilon^*), \max S_0(\epsilon^*), d]. \tag{13.25}$$

In der folgenden Tabelle wurde $K_0(\sqrt{c}s,\epsilon^*)$ für gegebenes s,ϵ^* mit dem Intervallhalbierungsverfahren ermittelt. $K_0^*(\epsilon^*)$ wurde durch das Maximum von $K_0(\sqrt{c}s,\epsilon^*)$ aus 100 äquidistanten Stützstellen aus $S_0(\epsilon^*)$ approximiert. Die Lösung d von (13.25) wurde mit dem Intervallhalbierungsverfahren ermittelt.

ϵ^*	$\min S_0(\epsilon^*)$	$\max S_0(\epsilon^*)$	$K_0^*(\epsilon^*)$	d
$c = 10$				
0.017	0.1	2.967	0.0926	7.66
0.01	0.1	1.639	0.0405	6.416
0.005	0.1	1.242	0.0177	6.118
0.004	0.1	1.185	0.0139	6.108
0.003	0.1	1.133	0.0103	6.124
0.001	0.1	1.041	0.0034	6.32
0.0003	0.1	1.012	0.001	6.614
$c = 4.383$				
0.04	0.228	2.952	0.1446	5.493
0.02	0.228	1.494	0.0687	4.686
0.015	0.228	1.33	0.0515	4.642
0.01	0.228	1.198	0.0344	4.661
0.002	0.228	1.034	0.0069	5.025

Mit analoger Rechnung könnte Satz 13.4 auch auf andere Mischungen zweier Normalverteilungen angewendet werden, d.h. $\epsilon \neq \frac{1}{2}$, andere und unterschiedliche Varianzen. Bei Mischungen von mehr als zwei Komponenten wird es etwas komplizierter.

Die höchstmögliche Abweichung $|\beta(g,Q) - a_1|$, die Satz 13.4 liefert, ist sehr klein. Die Schranken für $\sigma^2(g,Q)$ sind dagegen ziemlich großzügig. Offenbar ist der Satz auf weniger weit voneinander entfernte Mischungskomponenten anwendbar, wenn c kleiner gewählt wird. Das deckt sich mit meiner Erfahrung in der Anwendung auf Datensätze: Mit kleinerem c werden weniger gut voneinander getrennte Fixpunktcluster gefunden. Leider hat eine kleine Wahl von c den Nachteil, daß häufig extrem viele Cluster gefunden werden, so daß die Ausgabe des Verfahrens sehr unübersichtlich wird.

Es folgen einige Konsequenzen aus Satz 13.4. Ich gebe Bedingungen dafür an, daß ein KQ-FPCI g mit $\beta(g,Q) \in M_0(\epsilon^*) \to \{\beta_0\}$ existiert, d.h. der Lokationsparameter $\beta(g,Q)$ approximativ Fisher-konsistent ist. Es gelten die Bezeichungen von Satz 13.4, c sei fest, so daß (13.11) erfüllt ist. Außerdem sei

$$Q(\epsilon,\sigma_0^2) := (1-\epsilon)\mathcal{N}_{(\beta_0,\sigma_0^2)} + \epsilon H^*.$$

Falls σ_0^2 variiert, schreibe ich $M_0(\epsilon^*,\sigma_0^2)$ statt $M_0(\epsilon^*)$, entsprechend $S_0(\epsilon^*,\sigma_0^2)$ und $M_1(\epsilon^*,\sigma_0^2)$.

Hilfssatz 13.7 *Es gilt*

$$\epsilon^* \searrow 0 \Rightarrow M_0(\epsilon^*) \searrow \{\beta_0\}, \quad M_1(\epsilon^*) \searrow [\beta_0 - \sigma_0, \beta_0 + \sigma_0],$$
$$\sigma_0 \searrow 0 \Rightarrow M_0(\epsilon^*,\sigma_0^2) \searrow \{\beta_0\}, \quad M_1(\epsilon^*,\sigma_0^2) \searrow \{\beta_0\},$$
$$\epsilon^* \searrow 0 \Rightarrow S_0(\epsilon^*) \searrow \left[\frac{\sigma_0^2}{c}, \sigma_0\right].$$

Beweis: Die Grenzwertaussage für $S_0(\epsilon^*)$ folgt aus der Definition von $S_0(\epsilon^*)$. Für festes ϵ^* ist $K_0^*(\epsilon^*)$ endlich, also gelten die Aussagen für σ_0^2 nach Definition von $M_0(\epsilon^*)$, $M_1(\epsilon^*)$.

Weiter ist wegen der Monotonieeigenschaften von k

$$K_0^*(\epsilon^*) < k_0 \Leftrightarrow \sup_{s^2 \in S_0(\epsilon^*)} k(k_0, \sqrt{c}s, \epsilon^*) < 0.$$

Da k monoton in ϵ^* steigt, steigt auch K_0^* schwach monoton in ϵ^*, also $M_0(\epsilon_1^*) \supset M_0(\epsilon_2^*)$ für $\epsilon_1^* > \epsilon_2^*$. Mit $\sqrt{c}s \geq 1$ für $s^2 \in S_0(\epsilon^*)$ und Hilfssatz 11.6 gilt insbesondere die Inklusion

$$k_0 - E(k_0, 1) > \epsilon^* \sqrt{c} \max S_0(\epsilon^*) \Rightarrow \tag{13.26}$$
$$\Rightarrow \forall s^2 \in S_0(\epsilon^*): \; k(k_0, \sqrt{c}s, \epsilon^*) = E(k_0, \sqrt{c}s) - k_0 + \epsilon^* \sqrt{c}s < 0.$$

Mit Hilfssatz 11.4 steigt $k_0 - E(k_0, 1)$ streng monoton in k_0 gegen 1, ist stetig in k_0 und es ist $E(0, 1) = 0$. Außerdem konvergiert $\max S_0(\epsilon^*)$ für $\epsilon^* \to 0$ gegen σ_0^2. Für beliebig kleines $k_0 > 0$ existiert also ϵ_3^*, so daß (13.26) für alle $\epsilon^* < \epsilon_3^*$ erfüllt ist. Daher $K_0^*(\epsilon^*) \to 0$, $M_0(\epsilon^*) \to \{\beta_0\}$ für $\epsilon^* \to 0$.

Korollar 13.8 *Existiert mit $\tau > 0$ eine Umgebung $[\beta_0 - \sigma_0(1+\tau), \beta_0 + \sigma_0(1+\tau)]$ von β_0, auf der H^* stetig ist, dann*

$$\exists \epsilon_1 > 0 \quad \forall \epsilon \leq \epsilon_1 \quad \exists \epsilon^*(\epsilon), \; KQ\text{-}FPCI \; g \; bzgl. \; Q(\epsilon, \sigma_0^2):$$
$$\beta[g, Q(\epsilon, \sigma_0^2)] \in M_0[\epsilon^*(\epsilon)], \quad \sigma^2[g, Q(\epsilon, \sigma_0^2)] \in S_0[\epsilon^*(\epsilon)].$$
$$\epsilon \searrow 0 \Rightarrow M_0[\epsilon^*(\epsilon)] \searrow \{\beta_0\}, \quad S_0[\epsilon^*(\epsilon)] \searrow \left[\frac{\sigma_0^2}{c}, \sigma_0^2\right].$$

Beweis: Mit $P = \mathcal{N}_{(\beta_0, \sigma_0^2)}$ gilt

$$\inf_{m \in M_0(\epsilon^*), s^2 \in S_0(\epsilon^*)} P[m - \sqrt{c}s, m + \sqrt{c}s] = P[\max M_0(\epsilon^*) - \sigma_0, \max M_0(\epsilon^*) + \sigma_0] =: P_{\epsilon^*},$$

denn für festen Intervallmittelpunkt m wird $[m - \sqrt{c}s, m + \sqrt{c}s]$ offenbar durch minimales s am kürzesten und für festes s wird $P[m - \sqrt{c}s, m + \sqrt{c}s]$ durch maximales m wegen der Monotonieeigenschaften von φ und der Symmetrie von $M_0(\epsilon^*)$ um β_0 minimiert. Mit

$$\frac{H^*[m - \sqrt{c}s, m + \sqrt{c}s]}{P[m - \sqrt{c}s, m + \sqrt{c}s]} \leq \frac{1}{P_{\epsilon^*}} = \epsilon_0$$

ist offenbar (13.12) erfüllt. $\epsilon^* = \frac{\epsilon\epsilon_0}{1-\epsilon}$ konvergiert mit ϵ gegen 0 und ist für genügend kleines ϵ so klein, daß (13.11) erfüllt ist. Mit Hilfssatz 13.7 gilt irgendwann

$$M_1(\epsilon^*) \subset [\beta_0 - \sigma_0(1+\tau), \beta_0 + \sigma_0(1+\tau)].$$

Dann ist H^* auch stetig auf $M_1(\epsilon^*)$ und alles folgt aus Satz 13.4 und Hilfssatz 13.7.

Korollar 13.9 *Ist für festes ϵ, ϵ^* und gegebenes σ_1^2 sogar*

$$\forall[m-\sqrt{cs}, m+\sqrt{cs}],\ m \in M_0(\epsilon^*, \sigma_1^2), s^2 \in (0, \max S_0(\epsilon^*, \sigma_1^2)]$$

(13.12) erfüllt, dann

$$\forall \sigma_2^2 \leq \sigma_1^2\ \exists\ KQ\text{-}FPCI\ g\ bzgl.\ Q(\epsilon, \sigma_2^2):$$
$$\beta[g, Q(\epsilon, \sigma_2^2)] \in M_0(\epsilon^*, \sigma_2^2),\ \sigma^2[g, Q(\epsilon, \sigma_0^2)] \in S_0(\epsilon^*, \sigma_2^2).$$
$$\sigma_2^2 \searrow 0 \Rightarrow M_0(\epsilon^*, \sigma_2^2) \searrow \{\beta_0\}.$$

Für die Existenz eines solchen σ_2^2 ist hinreichend, daß H^* in einer Umgebung von β_0 eine beschränkte λ–Dichte hat.

Beweis: Nach den Voraussetzungen ist für ϵ_0, σ_1^2 Satz 13.4 anwendbar, denn H^* ist auch stetig auf $M_1(\epsilon^*, \sigma_1^2)$: Für jedes $a_0 \in M_1(\epsilon^*)$ (außer $a_0 = \sup M_1(\epsilon^*)$; dann ersetze man unten das „+" durch ein „−") gilt mit (13.12):

$$H^*\{a_0\} \leq \lim_{\tau \to 0} H^*[a_0, a_0+\tau] \leq \epsilon_0 \lim_{\tau \to 0} P[a_0, a_0+\tau] = 0,$$

da für τ hinreichend klein $a_0 + \frac{\tau}{2} \in M_0(\epsilon^*)$ und $\frac{\tau}{2} \leq \sqrt{\max S_0(\epsilon^*, \sigma_1^2)}$, so daß $[a_0, a_0+\tau]$ die Form $[m-\sqrt{cs}, m+\sqrt{cs}], m \in M_0(\epsilon^*), s^2 \leq \max S_0(\epsilon^*, \sigma_1^2)$ hat.

Für $\sigma_2^2 < \sigma_1^2$ ist $S_0(\epsilon^*, \sigma_2^2) \subset (0, \max S_0(\epsilon^*, \sigma_1^2)]$ und mit Hilfssatz 13.7 damit auch $M_1(\epsilon^*, \sigma_2^2) \subset M_1(\epsilon^*, \sigma_1^2)$. Die Voraussetzungen für Satz 13.4 sind für σ_2^2 also für dieselben ϵ_0, ϵ^* erfüllt und $M_0(\epsilon^*, \sigma_2^2) \searrow \{\beta_0\}$ mit $\sigma_2^2 \searrow 0$ folgt aus Hilfssatz 13.7.

Hat H^* eine beschränkte λ–Dichte h in einer Umgebung von β_0, dann gilt für hinreichend kleines σ_2^2, weil $\lim_{\sigma_2^2 \to 0} M_1(\epsilon^*, \sigma_2^2) = \{\beta_0\}$,

$$h_{\epsilon^*, \sigma_2^2} := \sup_{y \in M_1(\epsilon^*, \sigma_2^2)} h(y) < \infty.$$

Damit

$$\inf_{m \in M_0(\epsilon^*, \sigma_2^2), s^2 \in (0, \max S_0(\epsilon^*, \sigma_2^2))} \frac{H^*[m-\sqrt{cs}, m+\sqrt{cs}]}{P[m-\sqrt{cs}, m+\sqrt{cs}]} \leq \frac{\sigma_2^2 h_{\epsilon^*, \sigma_2^2}}{\varphi\left(\frac{\max M_0(\epsilon^*, \sigma_2^2) - \beta_0}{\sigma_2^2}\right)} =: B.$$

Der Nenner von B ist $\varphi[K_0^*(\epsilon^*)]$ nach Definition von M_0 und damit unabhängig von σ_2^2. Also konvergiert B für $\sigma_2^2 \searrow 0$ gegen 0.

Seien $\epsilon_0, \epsilon^* = \frac{\epsilon\epsilon_0}{1-\epsilon}$ so gegeben, daß (13.11) erfüllt ist. Dann kann σ_2^2 so klein gewählt werden, daß $B \leq \epsilon_0$. Damit ist die Voraussetzung des Korollars erfüllt.

13.3 Überlappende Mischungen: Regression ohne Achsenabschnitt

In diesem Abschnitt wird ein zu Satz 13.4 analoges Resultat für die lineare Regression ohne Achsenabschnitt bewiesen. Das bedeutet, daß in diesem Abschnitt immer $x \in \mathbb{R}^p$ und keine der Komponenten von x nach δ_1 verteilt ist. Betrachtet wird erneut ein Modell

der Form $Q = (1-\epsilon)P + \epsilon H^*$. P ist nun wieder eine gemeinsame Verteilung von x, y, wobei die Regressoren x p-dimensional normalverteilt und der Störterm $y - x'\beta_0$ davon unabhängig nach $\mathcal{N}_{(0,\sigma_0^2)}$ verteilt sein sollen. H^* soll eine Verteilung auf $\mathbb{R}^{p+1}$ sein, für die die folgenden Voraussetzungen gelten, die später präzisiert werden:

$$H^*(L) \leq \epsilon_0 P(L), \qquad E_{H^*}\left(\|d(x)\|^2 | (x,y) \in L\right) < k$$

für alle Mengen L der Form $\{(y - x'\theta)^2 \leq cs^2\}$ mit θ aus einer Umgebung M_0 von β_0 und s^2 aus einer Umgebung S_0 von σ_0^2. $d(x)$ sei die standardisierte Entfernung der Regressoren x zu ihrem Erwartungswert unter P. ϵ_0 und k sind gewisse Konstanten.

Die erste Ungleichung besagt wie schon in Satz 13.4, daß P und H^* gut voneinander getrennt sein müssen. Die zweite Ungleichung bedeutet, daß die Regressoren unter H^* nicht zu weit von denen von P entfernt sein dürfen. Anderenfalls könnte H^* eine Art „Hebelwirkung" auf den zu P gehörigen Cluster haben. Dieses Problem entsteht durch die Verwendung des KQ-Regressionsschätzers, der bekanntlich nicht robust ist. Siehe dazu die Bemerkung am Ende des Beweises. Die obere Schranke k ist allerdings größer als 40, so daß diese Voraussetzung nur bei einer extrem „abgelegenen" Verteilung H^* problematisch ist.

Für den für die Anwendung interessanteren Fall der linearen Regression mit Achsenabschnitt sind die Ergebnisse komplementär: Satz 13.11 behandelt die Regression ohne Achsenabschnitt, Satz 13.4 behandelt eine Regression, die nur aus dem Achsenabschnitt besteht. Leider kann man durch Kombination dieser beiden Resultate nicht direkt ein Ergebnis für die Regression mit Achsenabschnitt gewinnen. Zwar kann man den KQ-Schätzer einer linearen Regression durch eine Lokationsschätzung und einen KQ-Schätzer einer Regression durch den Ursprung berechnen, wenn man die Daten geeignet transformiert. Es besteht aber kein einfacher Zusammenhang zwischen den Fixpunktclustern dieser beiden Probleme und denen des kombinierten Regressionsproblems mit Achsenabschnitt. Die Ergebnisse können also nur Indizien dafür sein, daß auch im Problem mit Achsenabschnitt entsprechende Fixpunktcluster existieren.

Wie im vorigen Abschnitt wird auch Satz 13.11 durch die Anwendung auf eine einfache Mischung von Normalverteilungs-Modellen illustriert werden (Beispiel 13.14). Erneut wird sich zeigen, daß die Theorie nur dann anwendbar ist, wenn die Mischungskomponenten sehr gut voneinander getrennt sind. Danach (Korollar 13.18 und 13.19) folgt wieder eine Untersuchung der Fälle $\epsilon \searrow 0$, $\sigma_0^2 \searrow 0$.

Bezeichnungen und Konstanten für Satz 13.11: Sei

$$P(x,y) = \int 1(u \leq x)\Phi_{(0,\sigma_0^2)}(y - u'\beta_0)d\mathcal{N}_{(0,AA')_p}(u) \tag{13.27}$$

wie im Mischmodell mit zufälligen Regressoren 3 (allerdings ohne Achsenabschnitt). Dabei sei A eine invertierbare $p \times p$-Matrix und $\sigma_0 \geq 0$.

Es werden Konstanten t, ϵ^* und eine Fixpunktcluster-Justierkonstante c benötigt, die folgende Ungleichungen erfüllen:

$$0 < t \leq \tfrac{1}{2}, \tag{13.28}$$

$$0 \leq \epsilon^* \leq \frac{0.04256t^2(1+0.8587t^2)}{4c(1+t^2)(1+0.5025t^2)+(c-1)0.04256t^2(1+0.8587t^2)}, \tag{13.29}$$

$$c \geq \left(2\frac{\varphi(0)}{\varphi(\sqrt{1+t^2})}+1\right)(1+\epsilon^*)(1+t^2)^2, \tag{13.30}$$

$$c_2(\epsilon^*) := 1-(c-1)\epsilon^* > 0. \tag{13.31}$$

Aus (13.30) folgt $c \geq \left(2\frac{\varphi(0)}{\varphi(1)}+1\right) = 4.2974$. Damit ergibt sich aus (13.29) mit $0 \leq t \leq \frac{1}{2}$:

$$\epsilon^* \leq \frac{0.04256t^2(1+0.8587t^2)}{4c} \leq 0.00075. \tag{13.32}$$

Also sind (13.30) und (13.31) immer erfüllt, wenn c gemäß

$$1+\frac{1}{0.00075} = 1331.01 > c \geq 1.00075(1.25)^2\left(2\frac{\varphi(0)}{\varphi(\sqrt{1.25})}+1\right) = 7.4065 \tag{13.33}$$

gewählt wird. Weiter gilt $c_2(\epsilon^*) < 1$.

Definiere für $(\theta, s^2) \in I\!R^p \times [0,\infty]$:

$$L(\theta, s^2) := \{(x,y) \in I\!R^{p-1} : (y-\theta' x)^2 \leq cs^2\}.$$

und weiter

$$S_0 := \left[\frac{\sigma_0^2(1+t^2)}{c}, \frac{\sigma_0^2(1+t^2)}{c_2(\epsilon^*)}\right],$$

$$M_0 := \left\{\theta : (\theta-\beta_0)' AA' (\theta-\beta_0) \in [0, \sigma_0^2 t^2]\right\},$$

$$M := M_0 \times S_0, \qquad I_0 := \bigcup_{s^2 \in S_0, \theta \in M_0} L(\theta, s^2).$$

Bemerkung 13.10 *Um die Voraussetzung (13.38) formulieren zu können, muß I_0 meßbar sein. Tatsächlich ist I_0 abgeschlossen und daher in $I\!B^{p+1}$.*

Beweis: Ich zeige, daß der Grenzwert jeder konvergenten Folge aus $I_0^{I\!N}$ in I_0 liegt. Damit ist I_0 abgeschlossen nach zum Beispiel Satz 155.7 aus Heuser (1981). Sei also

$$(x_n, y_n)_{n \in I\!N} \in I_0^{I\!N}, \ (x_n, y_n) \to_{n\to\infty} (x_0, y_0).$$

Für (x_n, y_n) gilt:

$$\exists (\theta_n, s_n^2) \in M : \ (y_n - \theta_n' x_n)^2 \leq cs_n^2.$$

Weiter gibt es, da M kompakt ist, eine Teilfolge

$$\left(\theta_{n_m}, s_{n_m}^2\right)_{m \in I\!N} \to_{m\to\infty} (\theta_0, s_0^2) \in M.$$

Wegen $(y_n - \theta_n' x_n)^2 - cs_n^2 \in [-c\max S_0, 0]$ kompakt, können die n_m so gewählt werden, daß

$$\lim_{m\to\infty}(y_{n_m} - \theta_{n_m}' x_{n_m})^2 - cs_{n_m}^2 = d_0 \in [-c\max S_0, 0].$$

Die Funktion

$$f : I\!R^{2p+1} \times I\!R^+ \mapsto I\!R, \ (x, y, \theta, s^2) \mapsto (y - x'\theta)^2 - cs^2$$

ist offenbar stetig. Also gilt

$$f(x_0, y_0, \theta_0, s_0^2) = \lim_{m\to\infty}(y_{n_m} - \theta_{n_m}' x_{n_m})^2 - cs_{n_m}^2 = d_0 \leq 0$$

und daher $(y_0 - \theta_0' x_0)^2 \leq s_0^2$, also $(x_0, y_0) \in I_0$, was zu zeigen war.

Satz 13.11 (Regression ohne Achsenabschnitt, überlappende Mischung) *Es gelten die eben eingeführten Bezeichnungen und Voraussetzungen. Ist nun*

$$Q(x,y) = (1-\epsilon)P(x,y) + \epsilon H^*(x,y), \tag{13.34}$$

$0 < \epsilon < 1$, *und* $H^* \in \mathcal{P}_{p+1}$ *erfülle*

$$\forall s^2 \in S_0, \theta \in M_0 :$$
$$H^*(L(\theta, s^2)) \leq \epsilon_0 P(L(\theta, s^2)), \tag{13.35}$$
$$H^*\{(y - x'\theta)^2 = cs^2\} = 0 \tag{13.36}$$
$$E_{H^*}\left(\|\mathbf{A}^{-1}x\|^2 \,|\, (x,y) \in L(\theta, s^2)\right) < \tfrac{1}{0.0847t^2}, \tag{13.37}$$
$$E_{H^*}\left(\|\mathbf{A}^{-1}x\|^2 \,|\, (x,y) \in I_0\right) < \infty, \tag{13.38}$$

wobei $\epsilon_0 := \frac{\epsilon^*(1-\epsilon)}{\epsilon}$;
dann $\exists$ *KQ-FPCI* g *bzgl.* Q *mit*

$$\beta(g,Q) \in M_0, \sigma^2(g,Q) \in S_0. \tag{13.39}$$

Beweis: Sei zunächst $\sigma_0^2 > 0$. Wie im Beweis von Satz 13.4 soll Brouwers Fixpunktsatz angewendet werden. Aufgrund der Äquivarianzen der KQ-FPCI (Bemerkung 8.5) sei ohne Einschränkung $\beta_0 = 0, \sigma_0^2 = 1, \mathbf{A} = \mathbf{I}_p$. Sei f definiert wie in Bemerkung 8.4,

$$M = M_0 \times S_0 = \{\theta : \|\theta\| \in [0,t]\} \times \left[\frac{1+t^2}{c}, \frac{1+t^2}{c_2(\epsilon^*)}\right].$$

Nach Schritt 1 und Bemerkung 8.4 genügt es zu zeigen, daß M einen Fixpunkt von f enthält. Schritt 2 und 3 bereiten Schritt 4 und 5 vor, die besagen, daß die Einschränkung von f auf M eine Selbstabbildung ist, die nach Schritt 6 stetig ist. Da M als Produkt kompakter und konvexer Mengen wieder kompakt und konvex und wegen $c > 1 > c_2(\epsilon^*), t > 0$ nichtleer ist, sichert erneut Brouwers Fixpunktsatz die Existenz eines Fixpunktes von f in M.

Sei $V(u,s)$ definiert wie in Abschnitt 11.2,

$$\beta(\theta, s^2) := \beta(g_{\theta,s^2}, Q), \qquad \sigma^2(\theta, s^2) := \sigma^2(g_{\theta,s^2}, Q), \tag{13.40}$$
$$V_{\theta_0,s}(\theta,\epsilon) := \frac{(1-\epsilon)\int (y-\theta' x)^2 g_{\theta_0,s^2}(x,y)dP(x,y) + \epsilon \int (y-\theta' x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y) + \epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)}, \tag{13.41}$$
$$V_{\theta_0,s}(\theta) := V_{\theta_0,s}(\theta, 0), \tag{13.42}$$
$$g_\theta := \frac{\sqrt{c}s}{\sqrt{1+\|\theta\|^2}}, \qquad k := 1 - V(0,1), \tag{13.43}$$
$$D_{(\theta_0,s)}(\theta_1,\theta_2) := V_{\theta_0,s}(\theta_1) - V_{\theta_0,s}(\theta_2). \tag{13.44}$$

Es werden folgende Rechnungen benötigt, die auch die Werte in (13.29) erklären:

$$k = 0.7089, \quad k^2 = 0.5025, \quad V(0,1) = 0.2911, \quad (1-k)^2 = V(0,1)^2 = 0.0847,$$
$$k^2 V(0,1)^2 = 0.04256, \quad (2 - V(0,1))k^2 = 0.8587. \tag{13.45}$$

Es folgen die Beweisschritte

Schritt 1: $(\theta, s^2) \in M \Rightarrow g_{\theta,s^2}$ erfüllt (8.3) und (8.4). Außerdem sind

$$\int \|x\|^2 g_{\theta,s^2} dQ(x,y) < \infty, \quad \int y^2 g_{\theta,s^2} dQ(x,y) < \infty.$$

Schritt 2:

$$s > 0 \Rightarrow V_{\theta_0,s}(\theta) = \frac{W(\theta,\theta_0)+(1+\theta'\theta_0)^2 V(0,s_{\theta_0})}{1+\|\theta_0\|^2}, \text{ wobei}$$
$$W(\theta,\theta_0) := \|\theta_0\|^2 + \|\theta\|^2 + \|\theta_0\|^2\|\theta\|^2 - (\theta'\theta_0)^2 - 2\theta'\theta_0 \geq 0. \tag{13.46}$$

Schritt 3: Aus Schritt 2 folgt unter $s > 0$:

$$\min_\theta V_{\theta_0,s}(\theta) \leq V_{\theta_0,s}(\theta_0) < 1 + \|\theta_0\|^2, \tag{13.47}$$
$$V_{\theta_0,s}(\theta) \geq \frac{V(0,s_{\theta_0})}{1+\|\theta_0\|^2}; \tag{13.48}$$
$$\|\theta\| > \|\theta_0\| \Rightarrow V_{\theta_0,s}(\theta) > V_{\theta_0,s}(\theta_0). \tag{13.49}$$

Schritt 4:

$$(\theta_0, s^2) \in M \Rightarrow \|\beta(\theta_0, s^2)\| \leq t.$$

Schritt 5:

$$(\theta_0, s^2) \in M \Rightarrow \sigma^2(\theta_0, s^2) \in S_0.$$

Schritt 6: f ist stetig auf M.

Schritt 7: Der Satz gilt auch für $\sigma_0^2 = 0$.

Beweis von Schritt 1: Sei $(\theta, s^2) \in M$, also auch $s^2 > 0$. Damit:

$$\int g_{\theta,s^2}(x,y) dQ(x,y) \geq (1-\epsilon) \int \int g_{\theta,s^2}(x,y) d\mathcal{N}_{(x'\beta_0,\sigma_0^2)}(y) d\mathcal{N}_{(0,I_p)}(x) > 0,$$

also erfüllt g_{θ,s^2} (8.3).

Nun zeige ich die Erfülltheit der Voraussetzungen von Hilfssatz 11.1, der (8.4) sichert, mit $dR = g_{\theta,s^2} dQ$.

Natürlich ist $\int \|x\|^2 dP(x,y) < \infty$ und damit auch $\int \|x^2\| g_{\theta,s^2}(x,y) dP(x,y) < \infty$. Nach Voraussetzung (13.38) ist dann schließlich auch

$$\int \|x^2\| g_{\theta,s^2}(x,y) d[(1-\epsilon)P + \epsilon H^*](x,y) < \infty.$$

Wegen $g_{\theta,s^2}(x,y) = 1[|y - x'\theta| \leq \sqrt{c}s]$ ist

$$\int y^2 g_{\theta,s^2}(x,y) dQ(x,y) \leq \int (x'\theta + \sqrt{c}s)^2 g_{\theta,s^2}(x,y) dQ(x,y).$$

$(x'\theta + \sqrt{c}s)^2$ und damit auch y^2 sind mit $\|x\|^2$ R–integrierbar.

Angenommen, $\int xx' g_{\theta,s^2}(x,y)dQ(x,y)$ wäre nicht invertierbar. Dann

$$\exists q \in I\!R^p : 0 = q' \int xx' g_{\theta,s^2}(x,y)dQ(x,y)q =$$
$$= q' \frac{\int xx' g_{\theta,s^2}(x,y)dQ(x,y)}{\int g_{\theta,s^2}(x,y)dQ(x,y)} q = E_Q[(q'x)^2 | g_{\theta,s^2}(x,y) = 1],$$
$$\text{also } Q[q'x = 0 | g_{\theta,s^2}(x,y) = 1] = 1.$$

Dann wäre aber

$$0 = \int 1(q'x \neq 0) g_{\theta,s^2}(x,y)dQ(x,y) \geq$$
$$\geq (1-\epsilon) \int 1(q'x \neq 0) g_{\theta,s^2}(x,y)dP(x,y) =$$
$$= (1-\epsilon) \int \int g_{\theta,s^2}(x,y)d\mathcal{N}_{(x'\beta_0,\sigma_0^2)}(y)d\mathcal{N}_{(0,I_p)}(x) > 0.$$

Also folgt (8.4) aus Hilfssatz 11.1.

Beweis von Schritt 2: Seien $\theta, \theta_0 \in I\!R^p$, $s > 0$,

$$r_{a,b}(x,y) := a(y - \theta_0'x) + b(y - \theta'x) \text{ für } a, b \in I\!R. \tag{13.50}$$

Sei für alle in diesem Beweisschritt auftauchenden E-Werte, Varianzen und Kovarianzen $\mathcal{L}((x',y)') = P = \mathcal{N}_{(0,I_{p+1})}$. Bezeichne für $h : I\!R^{p+1} \mapsto I\!R$

$$E_g(h(x,y)) := \frac{\int h(x,y) g_{\theta_0,s^2}(x,y)dP(x,y)}{\int g_{\theta_0,s^2}(x,y)dP(x,y)},$$

also den bedingten Erwartungswert von $h(x,y)$ unter $g_{\theta_0,s^2}(x,y) = 1$. Dann ist

$$V_{\theta_0,s}(\theta) = E_g\left[(y - \theta'x)^2\right].$$

Zur Berechnung von $V_{\theta_0,s}(\theta)$: Es gilt

$$\begin{pmatrix} -\theta_0' & 1 \\ -a\theta_0' - b\theta' & a+b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y - \theta_0'x \\ r_{a,b}(x,y) \end{pmatrix}. \tag{13.51}$$

Dieser Vektor ist also - falls $a \neq 0$ oder $b \neq 0$ - bivariat normalverteilt mit

$$\text{Cov}((y - \theta_0'x), r_{a,b}(x,y)) = a + b + (a\theta_0 + b\theta)'\theta_0. \tag{13.52}$$

Mit der Wahl $a := -(1 + \theta'\theta_0), b := 1 + \|\theta_0\|^2$ sind $(y - \theta_0'x)$ und $r_{a,b}(x,y)$ stochastisch unabhängig, da die Kovarianz dann 0 ist. Insbesondere sind $(y - \theta_0'x)$ und $r_{a,b}(x,y)$ dann auch bedingt unter $g_{\theta_0,s^2}(x,y) = 1((y - \theta_0'x)^2 \leq cs^2) = 1$ stochastisch unabhängig, was zunächst gezeigt wird: Sei $r_0 := y - \theta_0'x$, $f(r_0) := g_{\theta_0,s^2}(x,y)$ (letzteres hängt von (x,y) nur durch r_0 ab), seien $A, B \in I\!B$. Dann gilt:

$$P[r_0 \in A, r_{a,b}(x,y) \in B | f(r_0) = 1] = \frac{\int 1(r_0 \in A) 1(r_{a,b}(x,y) \in B) f(r_0) dP(x,y)}{\int f(r_0)dP(x,y)} =$$
$$= \frac{\int 1(r_0 \in A) f(r_0)dP(x,y)}{\int f(r_0)dP(x,y)} \int 1(r_{a,b}(x,y) \in B)dP(x,y) =$$
$$= P[r_0 \in A | f(r_0) = 1] P[r_{a,b}(x,y) \in B] =$$
$$= P[r_0 \in A | f(r_0) = 1] P[r_{a,b}(x,y) \in B | f(r_0) = 1],$$

da $r_{a,b}(x,y)$ und r_0 stochastisch unabhängig unter P sind. Also folgt die Unabhängigkeit von $r_{a,b}(x,y)$ und r_0 bedingt unter $f(r_0)=1$.

Weiter ergibt sich aus (13.50):

$$E\left[r_{a,b}(x,y)^2\right] = a^2(1+\|\theta_0\|^2) + 2ab(1+\theta'\theta_0) + b^2(1+\|\theta\|^2) = \\ = (1+\|\theta_0\|^2)\left[(1+\|\theta_0\|^2)(1+\|\theta\|^2) - (1+\theta'\theta_0)^2\right]. \tag{13.53}$$

Zur Berechnung von $V_{\theta_0,s}(\theta)$ wird die Unabhängigkeit von $r_{a,b}(x,y)$ und r_0 (und also von $r_{a,b}(x,y)$ und g_{θ_0,s^2}) benutzt:

$$E\left[r_{a,b}(x,y)^2\right] = E_g\left[r_{a,b}(x,y)^2\right] = \\ = a^2 V_{\theta_0,s}(\theta_0) + 2abE_g[(y-\theta_0'x)(y-\theta'x)] + b^2 V_{\theta_0,s}(\theta). \tag{13.54}$$

Weiter ist

$$0 = E[r_{a,b}(x,y)] = E_g[r_{a,b}(x,y)] = E_g[r_{a,b}(x,y)]E_g(y-\theta_0'x) = \\ = E_g[a(y-\theta_0'x)^2 + b(y-\theta'x)(y-\theta_0'x)] \Rightarrow \\ \Rightarrow bE_g[(y-\theta_0'x)(y-\theta'x)] = -aV_{\theta_0,s}(\theta_0).$$

Einsetzen in (13.54) und auflösen nach $V_{\theta_0,s}(\theta)$:

$$V_{\theta_0,s}(\theta) = \frac{E\left[r_{a,b}(x,y)^2\right] + a^2 V_{\theta_0,s}(\theta_0)}{b^2}. \tag{13.55}$$

Nun fehlt noch

$$V_{\theta_0,s}(\theta_0) = \frac{\int (y-\theta_0'x)^2 1((y-\theta_0'x)^2 < cs^2) dP(x,y)}{\int 1((y-\theta_0'x)^2 < cs^2) dP(x,y)} = \\ = \frac{\int u^2 1(u^2 < cs_0^2) d\mathcal{N}_{(0,1+\|\theta_0\|^2)}(u)}{\int 1(u^2 < cs^2) d\mathcal{N}_{(0,1+\|\theta_0\|^2)}(u)} = (1+\|\theta_0\|^2) V(0, s_{\theta_0}) \tag{13.56}$$

(letztere Gleichung durch Substitution $t = \frac{u}{\sqrt{1+\|\theta_0\|^2}}$), und aus (13.53) ergibt sich

$$E\left[r_{a,b}(x,y)^2\right] = (1+\|\theta_0\|^2) W(\theta,\theta_0),$$

also (13.55) $\Leftrightarrow$ (13.46). Zuletzt bringt die Cauchy-Schwarz-Ungleichung noch

$$\|\theta\|^2\|\theta_0\|^2 \geq (\theta'\theta_0)^2, \text{ also} \\ W(\theta,\theta_0) \geq \|\theta_0\|^2 + \|\theta\|^2 - 2\theta'\theta_0 \geq \\ \geq \|\theta_0\|^2 + \|\theta\|^2 - 2\|\theta_0\|\|\theta\| = (\|\theta\| - \|\theta_0\|)^2 \geq 0. \tag{13.57}$$

Beweis von Schritt 3: Für alle Teile dieses Beweisschrittes wird $V(0, s_{\theta_0}) < 1$ gebraucht, was aus Hilfssatz 11.7 folgt.

Zuerst folgt daraus (13.47), denn

$$\min_\theta V_{\theta_0,s}(\theta) \leq V_{\theta_0,s}(\theta_0) = (1+\|\theta_0\|^2) V(0, s_{\theta_0}), \tag{13.58}$$

wobei die letzte Gleichung bereits als (13.56) gezeigt wurde.

Um (13.48) zu zeigen, beweise ich zuerst

$$\theta'\theta_0 < 0 \Rightarrow V_{\theta_0,s}(-\theta) < V_{\theta_0,s}(\theta), \text{ also} \tag{13.59}$$

$$\theta \text{ minimiert } V_{\theta_0,s}(\theta) \Rightarrow \theta'\theta_0 \geq 0. \tag{13.60}$$

Beweis von (13.59): Sei $\theta'\theta_0 < 0$. Dann errechnet man mit Schritt 2

$$\begin{aligned}(1+\|\theta_0\|^2)[V_{\theta_0,s}(-\theta) - V_{\theta_0,s}(\theta)] = \\ = W(-\theta,\theta_0) - W(\theta,\theta_0) + [(1-\theta'\theta_0)^2 - (1+\theta'\theta_0)^2]V(0,s_{\theta_0}) = \\ = 4\theta'\theta_0(1 - V(0,s_{\theta_0})) < 0.\end{aligned}$$

Beweis von (13.48): Mit $\theta_1 := \arg\min_\theta V_{\theta_0,s}(\theta)$ (irgendeines, falls es mehrere gibt) gilt für beliebiges θ mit Hilfe von Schritt 2 und (13.60):

$$V_{\theta_0,s}(\theta) \geq V_{\theta_0,s}(\theta_1) \geq \frac{(1+\theta_1'\theta_0)^2 V(0,s_{\theta_0})}{1+\|\theta_0\|^2} \geq \frac{V(0,s_{\theta_0})}{1+\|\theta_0\|^2}.$$

Nun zeige ich noch (13.49). Es gilt $W(\theta_0,\theta_0) = 0$, so daß Schritt 2

$$V_{\theta_0,s}(\theta_0) - V_{\theta_0,s}(\theta) = \frac{-W(\theta,\theta_0) + [(1+\|\theta_0\|^2)^2 - (1+\theta'\theta_0)^2]V(0,s_{\theta_0})}{1+\|\theta_0\|^2} \tag{13.61}$$

bringt. Für $\|\theta_0\| < \|\theta\|$ ist das kleiner als 0: Falls $[(1+\|\theta_0\|^2)^2 - (1+\theta'\theta_0)^2] \geq 0$, dann folgt für den Zähler von (13.61) mit Hilfe von $V(0,s_{\theta_0}) < 1$:

$$\begin{aligned}[(1+\|\theta_0\|^2)^2 - (1+\theta'\theta_0)^2]V(0,s_{\theta_0}) - W(\theta,\theta_0) \leq \\ \leq [(1+\|\theta_0\|^2)^2 - (1+\theta'\theta_0)^2] - W(\theta,\theta_0) = \\ = 2\|\theta_0\|^2 + \|\theta_0\|^4 - 2\theta'\theta_0 - (\theta'\theta_0)^2 - W(\theta,\theta_0) = (1+\|\theta_0\|^2)(\|\theta_0\|^2 - \|\theta\|^2) < 0.\end{aligned}$$

Anderenfalls ist mit $[(1+\|\theta_0\|^2)^2 - (1+\theta'\theta_0)^2]$ auch der Zähler von (13.61) und damit der ganze Ausdruck kleiner als 0.

Beweis von Schritt 4: Teil 1: Wir betrachten zuerst den Fall $kt \leq \|\theta_0\| \leq t$, k definiert in (13.43). Gezeigt wird:

$$\|\theta\| > \|\theta_0\| \Rightarrow V_{\theta_0,s}(\theta,\epsilon) > V_{\theta_0,s}(k\theta_0,\epsilon). \tag{13.62}$$

$\|\beta(\theta_0, s^2)\| \leq t$ folgt aus (13.62) nach Definition (8.6), denn

$$\beta(\theta_0, s^2) = \arg\min_{\theta_1} V_{\theta_0,s}(\vartheta_1,\epsilon).$$

Um (13.62) zu beweisen, benötigt man im Falle $\int g_{\theta_0,s^2}(x,y)dH^*(x,y) > 0$:

$$\frac{\int (y-k\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} < 4cs^2 \text{ und} \tag{13.63}$$

$$D_{(\theta_0,s)}(\theta, k\theta_0) \geq 4cs^2\epsilon^*. \tag{13.64}$$

Beweis von (13.63):

$$\frac{\int(y-k\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} = \frac{\int[y-\theta_0'x-(k\theta_0-\theta_0)'x]^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} =$$

$$= \frac{\int(y-\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} - 2\frac{\int(y-\theta_0'x)(k\theta_0-\theta_0)'x\, g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} +$$

$$+\frac{\int((k\theta_0-\theta_0)'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} \leq$$

(Der erste Term ist aufgrund der Definition von g_{θ_0,s^2} nicht größer als cs^2, der dritte Term wird mit der Cauchy-Schwarz-Ungleichung abgeschätzt:)

$$\leq cs^2 + 2(1-k)\theta_0'\frac{\int x(y-x'\theta_0)g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} +$$

$$+(1-k)^2\|\theta_0\|^2\frac{\int\|x\|^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} \leq$$

(Es gilt $g_{\theta_0,s^2}(x,y) = 1 \Rightarrow |y-x'\theta_0| < \sqrt{cs}$ und $1-k > 0$. Damit kann auch der zweite Term abgeschätzt werden:)

$$\leq cs^2 + 2(1-k)\sqrt{cs}\|\theta_0\|\frac{\int\|x\|g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} +$$

$$+(1-k)^2\|\theta_0\|^2\frac{\int\|x\|^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} \leq$$

(Nun wende ich $cs^2 > \sqrt{cs} > 1$ an. Das gilt wegen $s^2 \in S_0$. Die letzte Abschätzung wird nachher gezeigt.)

$$\leq \left(1 + 2(1-k)\|\theta_0\|\frac{\int\|x\|g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} +\right.$$

$$\left. + (1-k)^2\|\theta_0\|^2\frac{\int\|x\|^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)}\right)cs^2 < 4cs^2. \qquad (13.65)$$

Zur letzten Abschätzung: (13.37) besagt für den dritten Term:

$$(1-k)^2\|\theta_0\|^2\frac{\int\|x\|^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} < 1, \qquad (13.66)$$

da $M_0 \ni \theta_0$, also $\|\theta_0\|^2 \leq t$ und $(1-k)^2 = 0.0847$. Die Abschätzung (13.65) folgt, weil damit und mit Ljapunovs Ungleichung auch der zweite Term abgeschätzt werden kann:

$$(1-k)\|\theta_0\|E_{H^*}[\|x\| \quad | g_{\theta_0,s^2}(x,y) = 1] \leq$$

$$\leq \sqrt{(1-k)^2\|\theta_0\|^2 E_{H^*}[\|x\|^2 \quad | g_{\theta_0,s^2}(x,y) = 1]} < 1.$$

Beweis von (13.64): Mit Schritt 2 und 3 errechnet man

$$D_{(\theta_0,s)}(\theta, k\theta_0) = V_{\theta_0,s}(\theta) - V_{\theta_0,s}(k\theta_0) > V_{\theta_0,s}(\theta_0) - V_{\theta_0,s}(k\theta_0) =$$

$$= \frac{-W(\theta_0,k\theta_0)+\left[(1+\|\theta_0\|^2)^2-(1+k\|\theta_0\|^2)^2\right]V(0,s_{\theta_0})}{1+\|\theta_0\|^2} =$$

$$= \frac{-(1+k^2-2k)\|\theta_0\|^2+\left[2(1-k)\|\theta_0\|^2+(1-k^2)\|\theta_0\|^4\right]V(0,s_{\theta_0})}{1+\|\theta_0\|^2} =$$

$$= \frac{(1-k)\|\theta_0\|^2[(2+(1+k)\|\theta_0\|^2)V(0,s_{\theta_0})-(1-k)]}{1+\|\theta_0\|^2} = Z(s_{\theta_0},\|\theta_0\|),$$

$$Z(a,b) := \frac{(1-k)b^2[(2+(1+k)b^2)V(0,a)-1+k]}{1+b^2}.$$

Wegen $s^2 \in S_0$ und $\|\theta_0\| \leq t$ ist $s_{\theta_0} \geq 1$. Mit Hilfssatz 11.7 steigt Z monoton in a. Setzt man 1 für s_{θ_0} ein, so steigt

$$Z(1,b) = \frac{(1-k)^2 b^2 [1+(1+k)b^2]}{1+b^2}$$

mit $\frac{b^2}{1+b^2}$ auch in $b = \|\theta_0\| \geq 0$, wobei $\|\theta_0\| \geq kt = (1-V(0,1))t$ vorausgesetzt war. Also erhalten wir

$$\begin{aligned} D_{(\theta_0,s)}(\theta_0, k\theta_0) \geq Z(1,kt) = \\ = \frac{(1-k)k^2t^2[2+(1+k)k^2t^2](1-k)+(1-k)}{1+k^2t^2} = \\ = \frac{t^2(1-V(0,1))^2V(0,1)^2\left[1+(2-V(0,1))(1-V(0,1))^2t^2\right]}{1+(1-V(0,1))^2t^2} =: D_1. \end{aligned} \tag{13.67}$$

Wie ich gleich zeigen werde, gilt

$$(13.29) \Rightarrow D_1 \geq 4cs^2\epsilon^*, \tag{13.68}$$

also (13.64).

Beweis von (13.63): Sei $B := \frac{D_1}{4c(1+t^2)}$, also mit (13.45)

$$B = \frac{0.04256t^2(1+0.8587t^2)}{4c(1+t^2)(1+0.5025t^2)}.$$

Dann gilt mit $\sup S_0 = \frac{1+t^2}{1-(c-1)\epsilon^*}$:

$$\begin{gathered} (13.29) \Leftrightarrow \epsilon^* \leq \frac{B}{1+(c-1)B} \Leftrightarrow \\ \Leftrightarrow [1+(c-1)B]\epsilon^* \leq B \Leftrightarrow \epsilon^* \leq [1-(c-1)\epsilon^*]B = \frac{D_1}{4c\sup S_0} \Leftrightarrow \\ \Leftrightarrow \epsilon^* 4c \sup S_0 \leq D_1. \end{gathered}$$

also (13.68) für $s^2 \in S_0$.

(13.62) folgt jetzt mit (13.63) und (13.64):

$$\begin{aligned} V_{\theta_0,s}(\theta,\epsilon) - V_{\theta_0,s}(k\theta_0,\epsilon) = \\ = \frac{(1-\epsilon)\left[\int (y-\theta'x)^2 g_{\theta_0,s^2}(x,y)dP(x,y) - \int (y-k\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dP(x,y)\right]}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)+\epsilon\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} + \\ + \frac{\epsilon\left[\int (y-\theta'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y) - \int (y-k\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)\right]}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)+\epsilon\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} \geq \end{aligned}$$

(Kürzen durch $(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)$. Im Falle $\int g_{\theta_0,s^2}(x,y)dH^*(x,y) = 0$ sind alle Terme mit H^* null, anderenfalls abschätzen: $\int (y-k\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)$ durch (13.63), $\int (y-\theta'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)$ durch 0:)

$$\geq \frac{D_{(\theta_0,s)}(\theta,k\theta_0) - 4cs^2\frac{\epsilon\int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}}{1+\frac{\epsilon\int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}} \geq$$

(Der Beweis dieser Ungleichung wird weiter unten nachgeliefert; die Argumentation ist ähnlich wie in Schritt 2 im Beweis von Satz 13.4.)

$$\geq \frac{D_{(\theta_0,s)}(\theta, k\theta_0) - 4cs^2\epsilon^*}{1+\epsilon^*}. \tag{13.69}$$

Zusammengesetzt mit (13.64) ergibt sich

$$V_{\theta_0,s}(\theta,\epsilon) - V_{\theta_0,s}(k\theta_0,\epsilon) \geq \frac{D_{(\theta_0,s)}(\theta,k\theta_0) - 4cs^2\epsilon^*}{1+\epsilon^*} > 0, \tag{13.70}$$

also (13.62) im Falle $kt \leq \|\theta_0\| \leq t$. Noch nachzutragen ist der Beweis von (13.69) (letzte Ungleichung): Seien $d > 0, h, b, a \geq 0, e \geq b$. Dann gilt:

$$\frac{a-bh}{d+b} = \frac{ad - bhe + ae - bhd}{(d+b)(d+e)} \geq \frac{ad - bhe + ab - ehd}{(d+b)(d+e)} = \frac{a-eh}{d+e}.$$

Die Ungleichung (13.69) folgt nun mit

$$a := D_{(\theta_0,s)}(\theta, k\theta_0), \qquad b := \frac{\int g_{\theta_{\Gamma},s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dP(x,y)},$$
$$d := 1, \qquad e := \epsilon_0, \qquad h := 4cs^2,$$

wobei (13.35) besagt, daß $e \geq b$. $a \geq 0$ gilt wegen (13.64), $d > 0, h, b \geq 0$ sind klar.

Teil 2: Nun sei noch $\|\theta_0\| < kt$. Gezeigt werden muß wieder $\|\beta(\theta_0,s)\| \leq t$, was diesmal aus

$$\|\theta\| \geq t \Rightarrow V_{\theta_0,s}(\theta,\epsilon) - V_{\theta_0,s}(\theta_0,\epsilon) > 0 \tag{13.71}$$

folgt. Sei also $\|\theta\| \geq t$. Um (13.71) zu zeigen, wird

$$D_{(\theta_0,s)}(\theta,\theta_0) \geq 4\epsilon^* cs^2 \tag{13.72}$$

benötigt.

Beweis von (13.72):

Fall 1: $\theta'\theta_0 \geq \|\theta_0\|^2$. Daraus folgt mit Schritt 2

$$D_{(\theta_0,s)}(\theta,\theta_0) = \frac{W(\theta,\theta_0) + [(1+\theta'\theta_0)^2 - (1+\|\theta_0\|^2)^2]V(0,s_{\theta_0})}{1+\|\theta_0\|^2} \geq \frac{W(\theta,\theta_0)}{1+\|\theta_0\|^2}.$$

(13.57) bringt dann

$$D_{(\theta_0,s)}(\theta,\theta_0) \geq \frac{(\|\theta\| - \|\theta_0\|)^2}{1+\|\theta_0\|^2} > \frac{(1-k)^2t^2}{1+k^2t^2} = \frac{t^2V(0,1)^2}{1+(1-V(0,1))^2t^2}.$$

Das ist aber größer als D_1 aus (13.67), denn

$$(1-V(0,1))^2\left[1+(2-V(0,1))(1-V(0,1))^2t^2\right] = 0.6104 < 1.$$

(13.68) bringt dann $D_{(\theta_0,s)}(\theta,\theta_0) \geq 4\epsilon^* cs^2$.

Fall 2: $0 \leq \theta'\theta_0 < \|\theta_0\|^2$, also $(1+\theta'\theta_0)^2 - (1+\|\theta_0\|^2)^2 < 0$. Wegen $V(0, s_{\theta_0}) < 1$ (Hilfssatz 11.7) gilt

$$D_{(\theta_0,s)}(\theta,\theta_0) \geq \frac{W(\theta,\theta_0)+[(1+\theta'\theta_0)^2-(1+\|\theta_0\|^2)^2]}{1+\|\theta_0\|^2} =$$
$$= \frac{\|\theta\|^2+\|\theta\|^2\|\theta_0\|^2-\|\theta_0\|^2-\|\theta_0\|^4}{1+\|\theta_0\|^2} = \|\theta\|^2 - \|\theta_0\|^2 \geq$$
$$\geq (1-k)^2t^2 > \frac{(1-k)^2t^2}{1+k^2t^2} \geq 4\epsilon^* cs^2$$

wie in Fall 1.

Fall 3: $\theta'\theta_0 < 0$, also mit (13.59)

$$V_{\theta_0,s}(\theta) > V_{\theta_0,s}(-\theta), \text{ also } D_{(\theta_0,s)}(\theta,\theta_0) > D_{(\theta_0,s)}(-\theta,\theta_0),$$

und Fall 1 oder 2 kann auf $-\theta$ angewendet werden.

Beweis von (13.71): Die folgende Rechnung ist identisch zu der Umformung, die zu (13.69) führt. Als einziger Unterschied kann statt (13.63) diesmal sogar

$$\frac{\int (y-\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{\int g_{\theta_0,s^2}(x,y)dH^*(x,y)} = E_{H^*}[(y-\theta_0'x)^2 | (y-\theta_0'x)^2 \leq cs^2] \leq cs^2 \quad (13.73)$$

abgeschätzt werden.

$$V_{\theta_0,s}(\theta,\epsilon) - V_{\theta_0,s}(\theta_0,\epsilon) \geq$$
$$\geq \frac{D_{(\theta_0,s)}(\theta,\theta_0) - cs^2 \frac{\epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}}{1+\frac{\epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}} \geq$$
$$\geq \frac{D_{(\theta_0,s)}(\theta,\theta_0) - \epsilon^* cs^2}{1+\epsilon^*} =: D_2$$

Aus (13.72) folgt $D_2 > 0$, also (13.71).

Beweis von Schritt 5: Sei $\|\theta_0\| \leq t, s^2 \in S_0$. Es soll $\sigma^2(\theta_0, s^2) = \min_{\theta_1} V_{\theta_0,s}(\theta_1, \epsilon)$ abgeschätzt werden. Zunächst gilt

$$\sigma^2(\theta_0, s^2) \leq V_{\theta_0,s}(\theta_0,\epsilon) \leq \frac{1+\|\theta_0\|^2+\epsilon^* cs^2}{1+\epsilon^*}; \quad (13.74)$$

die zweite Ungleichung erhält man wie folgt:

$$V_{\theta_0,s}(\theta_0,\epsilon) =$$
$$= \frac{(1-\epsilon)\int (y-\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dP(x,y) + \epsilon \int (y-\theta_0'x)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y) + \epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)} \leq$$

(Kürzen durch $(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)$ und abschätzen durch (13.73):)

$$\leq \frac{V_{\theta_0,s}(\theta_0) + cs^2 \frac{\epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}}{1+\frac{\epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}} \leq$$

(Die folgende Ungleichung folgt mit (13.20) aus dem Beweis von Satz 13.4; $a := V_{\theta_0,s}(\theta_0)$, $b := \frac{\epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}$, $d := 1$, $e := \epsilon^*$, $h := cs^2$. Die Voraussetzung $e \geq b$ folgt aus (13.35), weiter gilt $h \geq E_P[(y-\theta_0'x)^2 | (y-\theta_0'x)^2 \leq cs^2] = \frac{a}{d}$.)

$$\leq \frac{V_{\theta_0,s}(\theta_0) + \epsilon^* cs^2}{1+\epsilon^*} \leq \frac{1+\|\theta_0\|^2 + \epsilon^* cs^2}{1+\epsilon^*};$$

die letzte Ungleichung folgt mit (13.47), also gilt (13.74). Aus (13.74) ergibt sich die Implikation:

$$s^2 \leq \tfrac{1+t^2}{1+(1-c)\epsilon^*} \Rightarrow \sigma^2(\theta_0, s^2) \leq \tfrac{1+t^2+\epsilon^* c \frac{1+t^2}{1+(1-c)\epsilon^*}}{1+\epsilon^*} =$$
$$= \tfrac{(1+t^2)(1+(1-c)\epsilon^*)+\epsilon^* c(1+t^2)}{(1+\epsilon^*)(1+(1-c)\epsilon^*)} = \tfrac{1+t^2}{1+(1-c)\epsilon^*},$$

d.h. $\sigma^2(\theta_0, s^2)$ wird nicht größer als die Obergrenze von S_0.

Weiter gilt:

$$\sigma^2(\theta_0, s^2) = \min_\theta V_{\theta_0,s}(\theta, \epsilon) \geq$$

(Abschätzung $\int (y - x'\theta)^2 g_{\theta_0,s^2}(x,y)dH^*(x,y) \geq 0$)

$$\geq \min_\theta V_{\theta_0,s}(\theta) \frac{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)}{(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y) + \epsilon \int g_{\theta_0,s^2}(x,y)dH^*(x,y)} \geq$$

(Kürzen durch $(1-\epsilon)\int g_{\theta_0,s^2}(x,y)dP(x,y)$, abschätzen durch (13.35) und (13.48):)

$$\geq \frac{V(0, s_{\theta_0})}{(1+\epsilon^*)(1+\|\theta_0\|^2)} =: V_1(s^2).$$

Die Abschätzung von $\sigma^2(\theta_0, s^2)$ nach unten wird sich aus $V_1(\inf S_0) > \inf S_0$ ergeben, denn wegen Hilfssatz 11.7 ist $V(0,s)$ streng monoton wachsend in s, so daß für $s^2 > \inf S_0$ gilt:

$$\sigma^2(\theta_0, s^2) \geq V_1(s^2) > V_1(\inf S_0) > \inf S_0.$$

Dazu zeige ich $V_1(s^2) > s^2$ unter der Voraussetzung

$$0 < s_{\theta_0} \leq \sqrt{1+t^2}. \tag{13.75}$$

Wegen $\|\theta_0\| \leq t$ ist diese Voraussetzung für $s^2 \leq \inf S_0$ und insbesondere für $s^2 = \inf S_0$ erfüllt, so daß dann alles gezeigt ist.

Definiere $b := 1 + \|\theta_0\|^2$, also $s_{\theta_0} = \sqrt{\frac{c}{b}}s$, $h(s) := \frac{V(0,s_{\theta_0})}{b(1+\epsilon^*)} - s^2$ und

$$h_0(s) := \left[1 - b(1+\epsilon^*)s^2\right]\left[\Phi(s_{\theta_0}) - \Phi(-s_{\theta_0})\right] - 2s_{\theta_0}\varphi(s_{\theta_0}).$$

Mit (11.12) ist

$$h(s) = \tfrac{1}{b(1+\epsilon^*)}\left(1 - \tfrac{2s_{\theta_0}\varphi(s_{\theta_0})}{\Phi(s_{\theta_0})-\Phi(-s_{\theta_0})}\right) - s^2, \text{ also}$$
$$h(s) > 0 \Leftrightarrow h_0(s) > 0.$$

Mit $\frac{\partial}{\partial s} s_{\theta_0} = \sqrt{\frac{c}{b}}$ ist weiter

$$\begin{aligned} h_0'(s) &= -2sb(1+\epsilon^*)\left[\Phi(s_{\theta_0}) - \Phi(-s_{\theta_0})\right] + \\ &+ \left[1 - b(1+\epsilon^*)s^2\right] 2\sqrt{\tfrac{c}{b}}\varphi(s_{\theta_0}) - 2\sqrt{\tfrac{c}{b}}\varphi(s_{\theta_0})(1 - s_{\theta_0}^2) = \\ &= 2b(1+\epsilon^*)s\left(\sqrt{\tfrac{c}{b}}s\left(\tfrac{c}{b^2(1+\epsilon^*)} - 1\right)\varphi(s_{\theta_0}) - \left[\Phi(s_{\theta_0}) - \Phi(-s_{\theta_0})\right]\right) = \\ &= 2(1+\epsilon^*)\sqrt{bc}s^2\left(\left(\tfrac{c}{b^2(1+\epsilon^*)} - 1\right)\varphi(s_{\theta_0}) - \tfrac{\Phi(s_{\theta_0}) - \Phi(-s_{\theta_0})}{s_{\theta_0}}\right) =: V_2. \end{aligned}$$

V_2 ist größer als 0: Aufgrund von (13.30) und der Voraussetzung $\|\theta_0\| \leq t$, also $b^2 \leq (1+t^2)^2$, wird V_2 verkleinert, wenn $\left(2\frac{\varphi(0)}{\varphi(\sqrt{1+t^2})} + 1\right)(1+\epsilon^*)b^2$ für c eingesetzt wird. Dadurch folgt $V_2 > 0$ aus

$$2\frac{\varphi(0)\varphi(s_{\theta_0})}{\varphi(\sqrt{1+t^2})} - \frac{\Phi(s_{\theta_0}) - \Phi(-s_{\theta_0})}{s_{\theta_0}} > 0,$$

was aus (13.75), d.h. $\frac{\varphi(s_{\theta_0})}{\varphi(\sqrt{1+t^2})} > 1$, folgt.

Weiter gilt $s = 0 \Rightarrow s_{\theta_0} = 0 \Rightarrow h(0) = h_0(0) = 0$. Für $s > 0$ ist nun $V_2 > 0$, also $h_0'(s) > 0$, damit $h_0(s) > 0$ und schließlich auch $h(s) > 0$, was nach Definition $V_1(s^2) > s^2$ impliziert.

Beweis von Schritt 6: Schritt 6 folgt aus Hilfssatz 11.2. Dabei sind die Voraussetzungen (11.5), (11.7) und (11.8) wegen Schritt 1 erfüllt. Die Voraussetzungen (13.38) und (13.36) gelten für H^* und auch für die Normalverteilung P, also für Q. Damit ist auch (11.6) erfüllt.

Beweis von Schritt 7: Sei nun $\sigma_0^2 = 0$ und ohne Einschränkung $\beta_0 = 0$, also $P\{y = 0\} = 1$. Dann ist $M = \{(0,0)\}$. Es gilt $\int g_{0,0}(x,y)dQ(x,y) \geq 1 - \epsilon > 0$, also erfüllt $g_{0,0}$ (8.3). In Schritt 1 geht die Voraussetzung $\sigma_0^2 > 0$ nur in den Nachweis von (8.3) ein, mit derselben Argumentation wie dort erfüllt aber $g_{0,0}$ (8.4). Weiterhin gilt

$$\int (y - x'\theta)^2 1(y^2 = 0)dQ(x,y) = 0 \Leftrightarrow \theta = 0,$$

da für $\theta \neq 0$

$$\int (y - x'\theta)^2 1(y^2 = 0)dP(x,y) = \int (x'\theta)^2 d\mathcal{N}_{(0,AA')_p}(x) > 0,$$

so daß $M_0 \ni \beta(0,0) = 0 = \sigma^2(0,0) \in S_0$. Damit ist $g_{0,0}$ KQ-FPCI bzgl. Q.

Bemerkung 13.12 *Die Verteilung der Regressoren hat in der Verteilung P in diesem Fall immer den Erwartungswert 0. Das liegt daran, daß für $x \in \mathbb{R}^p$ mit $Ex = 0$ unter linearen Transformationen der Form $D : \mathbb{R}^{p+1} \mapsto \mathbb{R}^{p+1}$ aus Bemerkung 2.4 immer $E(\Gamma x) = 0$ gilt. Satz 13.11 wäre mit einigem formalen Aufwand aber auch auf Regressorenverteilungen der Form $\mathcal{N}_{(\eta,AA')}$ zu übertragen, indem das Regressionsproblem ohne Achsenabschnitt als Problem mit $(x,y) \in \mathbb{R}^p \times \{1\} \times \mathbb{R}$ mit $\beta_{p+1} \equiv 0$ formuliert würde. Die Menge $\mathcal{D}$ aus (2.2) müßte dann auf diejenigen Transformationen beschränkt werden, die $\beta_{p+1} \equiv 0$ erhalten und die Definition und der Äquivarianznachweis für KQ-FPCI müßten entsprechend modifiziert werden.*

Bemerkung 13.13 *Die Voraussetzung (13.37) deutet darauf hin, daß das KQ-Fixpunktcluster-Verfahren Schwierigkeiten haben könnte, Cluster zu finden, wenn gleichzeitig andere Cluster oder einzelne Punkte mit extrem entfernten Regressoren im Datensatz sind. (13.37) wird nur in Schritt 4 benutzt, um*

$$E_{H^*}\left[(y-x'k\theta_0)^2|g_{\theta_0,s^2}(x,y)=1\right] < 4cs^2$$

zu zeigen. Würde man das KQ-Regressionsfunktional durch ein Funktional ersetzen, das eine beschränkte Funktion der Residuen anstatt ihres Quadrates minimieren würde, so könnte dieser Erwartungswert ohne die zusätzliche Voraussetzung beschränkt werden.

Viele robuste Regressionsfunktionale sind auf diese Weise definiert, zum Beispiel MM- oder S-Schätzer (siehe Abschnitt 3.5.1). Siehe dazu auch Bemerkung 8.2.

Es folgt die Anwendung des Satzes 13.11. Ich werde hier ganz bestimmte Mischungen von Regressionsverteilungen mit normalverteilten Regressoren behandeln, nämlich „kreuzförmige" Modelle: Eine Komponente mit Steigung 1, die andere mit Steigung -1, Regressoren für beide Komponenten $\mathcal{N}(0,1)$-verteilt, gleiche Varianz σ_0^2 der Störterme, Anteil jeder Komponente am Gesamtmodell $\frac{1}{2}$. Die Punkte 1-98 des künstlichen Datensatzes aus Abschnitt 10 (Abbildung 8) könnten auf diese Weise erzeugt worden sein. Ich werde zeigen, daß für beide Komponenten je ein KQ-FPCI vorhanden ist, wenn σ_0 hinreichend klein ist.

Beispiel 13.14 *Falls* $c=10, p=1, \beta_1=1, \beta_2=-1$,

$$Q=\tfrac{1}{2}Q_1+\tfrac{1}{2}Q_2, \text{ wobei}$$
$$Q_i(x,y)=\int 1(z\le x)\Phi_{(0,\sigma_0^2)}(y-z'\beta_i)d\mathcal{N}(z), \quad i=1,2,$$
$$\sigma_0<0.0000323 \qquad (\text{bzw. } \sigma_0<0.0000503 \text{ für } c=7.407),$$

dann existieren KQ-FPCI $g_i, i=1,2$, *mit*

$$|\beta(g_i,P)-\beta_i|\le 0.2\sigma_0, \qquad 0.104\le\tfrac{\sigma^2(g_i,P)}{\sigma_0^2}\le 1.00037$$
$$(\text{bzw. } 0.14042\le\tfrac{\sigma^2(g_i,P)}{\sigma_0^2}\le 1.00036 \text{ für } c=7.407).$$

Beweis: Es gelten die Bezeichnungen aus Satz 13.11. Mit $\epsilon=\frac{1}{2}$ gilt $\epsilon^*=\epsilon_0$. Es sei ohne Einschränkung $P:=Q_1, H^*:=Q_2$. Es soll also die Existenz eines KQ-FPCI, der zu Q_1 gehört, gezeigt werden, Q_2 wird als Ausreißerverteilung behandelt. Die umgekehrte Situation ist völlig analog. Es sind nun also (vgl. (13.27)) $A=1, \beta_0=1$.

Sei ein $0\le t\le\frac{1}{2}$ vorgegeben. t gibt in Satz 13.11 die höchstmögliche Abweichung des Regressionsparameters von Q_1 zum zugehörigen KQ-FPCI an. Weiter sei ϵ^* gemäß Voraussetzung (13.29) so groß wie möglich gewählt:

$$\epsilon^*=\frac{0.04256t^2(1+0.8587t^2)}{4c(1+t^2)(1+0.5025t^2)+(c-1)0.04256t^2(1+0.8587t^2)}.$$

$c=10$ und $c=7.407$ erfüllen (13.33). Zur Anwendung von Satz 13.11: Voraussetzung (13.36) ist für Normalverteilungen erfüllt. Voraussetzung (13.38) ist erfüllt, weil $s^2>$

$0 \Rightarrow H^*\{L(\theta, s^2))\} > 0$, also auch $H^*(I_0) > 0$, und damit

$$E_{H^*}(x^2) = E_{\mathcal{N}}(x^2) = 1 \Rightarrow E_{H^*}(x^2|(x,y) \in I_0) < \infty.$$

Ich zeige nun:

- Voraussetzung (13.37) ist erfüllt für $\sigma_0^2 < 1$.
- Voraussetzung (13.35) ist erfüllt, wenn σ_0 hinreichend klein ist.

Damit können dann die Werte aus dem Beispiel numerisch berechnet werden.

Zu Voraussetzung (13.37): Sei $\mathcal{L}(x,y) = Q_2$, $\mathcal{L}(u,v) = \mathcal{N}_{(0,I_2)}$. Dann gilt für $b \in \mathbb{R}$:

$$\mathcal{L}(x, y - bx) = \mathcal{L}(u, -(1+b)u + \sigma_0 v), \tag{13.76}$$

denn $y - bx = y + x - (1+b)x$. Außerdem ist $\mathcal{L}(y+x) = \mathcal{N}_{(0,\sigma_0^2)}$ und x ist von $y+x$ stochastisch unabhängig unter Q_2.

Abzuschätzen ist nun $E_{H^*}[x^2|(y-bx)^2 \le cs^2]$ für $(b,s^2) \in M$. Es gilt mit (13.76):

$$E_{H^*}[x^2|(y-bx)^2 \le cs^2] = E_{\mathcal{N}_{(0,I_2)}}\left[u^2 \middle| \left(u - \tfrac{\sigma_0 v}{1+b}\right)^2 \le \tfrac{cs^2}{(1+b)^2}\right] =$$

$$= \int d\mathcal{N}(v) \frac{\int d\mathcal{N}(u) u^2 1\left(\left[u - \frac{\sigma_0 v}{1+b}\right]^2 \le \frac{cs^2}{(1+b)^2}\right)}{\int d\mathcal{N}(u) 1\left(\left[u - \frac{\sigma_0 v}{1+b}\right]^2 \le \frac{cs^2}{(1+b)^2}\right)} =$$

(mit den Bezeichnungen aus Abschnitt 11.2)

$$= \int \left[V\left(\frac{\sigma_0 v}{1+b}, \frac{\sqrt{c}s}{1+b}\right) + E\left(\frac{\sigma_0 v}{1+b}, \frac{\sqrt{c}s}{1+b}\right)^2\right] d\mathcal{N}(v) \le$$

(Für $(w,z) \in \mathbb{R} \times \mathbb{R}^+$ gilt nach Hilfssatz 11.7: $V(w,z) < 1$, nach Hilfssatz 11.5: $0 \le |E(w,z)| \le |w|$ mit $E(0,z) = 0$ sowie $E(-w,z) = -E(w,z)$ nach (11.11).)

$$\le 1 + \tfrac{\sigma_0^2}{(1+b)^2} \int v^2 d\mathcal{N}(v) = 1 + \tfrac{\sigma_0^2}{(1+b)^2} < 2 < \tfrac{1}{0.0847t^2}$$

für $\sigma_0^2 < 1$ (was in diesem Beispiel gilt) und $b > 0$. Auch das gilt, da $b \in M_0$, d.h. $|b-1| < \sigma_0 t$. Damit ist Voraussetzung (13.37) erfüllt.

Nun zu Voraussetzung (13.35). Zu zeigen ist: Für hinreichend kleines σ_0 gilt $\frac{Q_2(L)}{Q_1(L)} < \epsilon_0$ für alle $L(b,s^2)$ mit $(b,s^2) \in M$.

Mit (13.76) und der analogen Aussage für Q_1 ist

$$\frac{Q_2(L(b,s^2))}{Q_1(L(b,s^2))} = \frac{\int 1((-(1+b)u+\sigma_0 v)^2 < cs^2) d\mathcal{N}(u) d\mathcal{N}(v)}{\int 1(((1-b)u+\sigma_0 v)^2 < cs^2) d\mathcal{N}(u) d\mathcal{N}(v)} =$$

$$= \frac{\Phi\left(\frac{\sqrt{c}s}{\sqrt{(1+b)^2+\sigma_0^2}}\right) - \Phi\left(\frac{-\sqrt{c}s}{\sqrt{(1+b)^2+\sigma_0^2}}\right)}{\Phi\left(\frac{\sqrt{c}s}{\sqrt{(1-b)^2+\sigma_0^2}}\right) - \Phi\left(\frac{-\sqrt{c}s}{\sqrt{(1-b)^2+\sigma_0^2}}\right)} =: f(b,s,\sigma_0).$$

Gesucht ist nun $\sup_{(b,s^2)\in M} f(b,s,\sigma_0)$. Wegen $1-t\sigma_0 \le b \le 1+t\sigma_0$ ist $\sqrt{(1-b)^2+\sigma_0^2}$ maximal mit $|1-b| = t\sigma_0$ und $\sqrt{(1+b)^2+\sigma_0^2}$ ist minimal mit $b = 1-t\sigma_0$. Also

$$\sup_{(b,s^2)\in M} f(b,s,\sigma_0) = \sup_{s^2\in S_0} f(1-t\sigma_0,s,\sigma_0) =$$

$$= \sup_{s^2\in S_0} \frac{\Phi\left(\frac{\sqrt{c}s}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}}\right) - \Phi\left(\frac{-\sqrt{c}s}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}}\right)}{\Phi\left(\frac{\sqrt{c}s}{\sqrt{t^2+1}\sigma_0}\right) - \Phi\left(\frac{-\sqrt{c}s}{\sqrt{t^2+1}\sigma_0}\right)} =$$

$$= \sup_{s_0\in\left[1,\sqrt{\frac{c}{c_2(\epsilon^*)}}\right]} \frac{\Phi\left(\frac{\sqrt{1+t^2}\sigma_0 s_0}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}}\right) - \Phi\left(\frac{-\sqrt{1+t^2}\sigma_0 s_0}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}}\right)}{\Phi\left(\frac{\sqrt{1+t^2}s_0}{\sqrt{t^2+1}}\right) - \Phi\left(\frac{-\sqrt{1+t^2}s_0}{\sqrt{t^2+1}}\right)} =: \sup_{s_0\in S_1} f_0(s_0,\sigma_0).$$

$f_0(s_0,\sigma_0)$ ist stetig in $s_0 \in S_1$ und σ_0. Offensichtlich ist $f_0(s_0,0) = 0 \quad \forall s_0 \in S_1$. Weiterhin gilt

$$\frac{\partial}{\partial\sigma_0}\frac{\sqrt{1+t^2}\sigma_0 s_0}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}} =$$

$$= \frac{\sqrt{1+t^2}s_0}{4-4t\sigma_0+(t^2+1)\sigma_0^2}\left(\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2} + \sigma_0\frac{4t-2\sigma_0(t^2+1)}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}}\right) > 0$$

für $\sigma_0 \le \frac{1}{2}, t \le \frac{1}{2}$, denn dann ist $4-4t\sigma_0+(t^2+1)\sigma_0^2 \ge 3$ und $\frac{2\sigma_0(t^2+1)}{\sqrt{4-4t\sigma_0+(t^2+1)\sigma_0^2}} < \frac{5}{4\sqrt{3}} < \sqrt{3}$. Das bedeutet, daß $f_0(s_0,\sigma_0)$ für beliebiges festes $s_0 \in S_1$ monoton gegen 0 fällt, falls $\sigma_0 \searrow 0$. Also fällt auch $\sup_{s_0\in S_1} f_0(s_0,\sigma_0)$ schwach monoton. Angenommen, es gälte nicht

$$\lim_{\sigma_0\to 0}\sup_{s_0\in S_1} f_0(s_0,\sigma_0) = 0. \tag{13.77}$$

Dann gäbe es aufgrund der Kompaktheit von $S_1 \times \left[0,\frac{1}{2}\right]$ ein $d \in S_1$ und eine Folge $(s_n,\sigma_n) \to_{n\to\infty} (d,0)$, so daß

$$\lim_{n\to\infty} f_0(s_n,\sigma_n) = h > 0, \text{ also } f_0(d,0) = h \neq 0$$

wegen der Stetigkeit von f_0, was nicht sein kann. Also gilt (13.77).

Insbesondere wird $\sup f_0$ damit natürlich auch bei hinreichend kleinem σ_0 kleiner als $\epsilon^* > 0$, womit die Voraussetzung (13.35) des Satzes 13.11 erfüllt ist.

Um die konkreten Werte zu berechnen, habe ich jeweis für gegebenes t, c, ϵ^* durch Intervallhalbierungsiteration den Wert σ_0 berechnet, für den gilt:

$$\sup_{s_0\in\left[1,\sqrt{\frac{c}{c_2(\epsilon^*)}}\right]} f_0(s_0,\sigma_0) = \epsilon^*.$$

Dabei habe ich für gegebenes σ_0 das Supremum von $f_0(s_0,\sigma_0)$ über s_0 durch den höchsten Wert an Stützstellen im Abstand 0.01 approximiert. Man erhält folgende Werte:

t	ϵ^*	$\frac{\min S_0}{\sigma_0^2}$	$\frac{\max S_0}{\sigma_0^2}$	σ_0
$c = 10$				
0.5	0.0002293	0.1250	1.0021	0.0001623
0.2	0.0000415	0.1040	1.0004	0.0000323
0.1	0.0000106	0.1010	1.0001	0.0000083
0.05	0.0000027	0.1003	1.0000	0.0000021
$c = 7.407$				
0.5	0.0003094	0.1688	1.0020	0.0002534
0.2	0.0000560	0.1404	1.0004	0.0000503
0.1	0.0000143	0.1364	1.0001	0.0000131
0.05	0.0000036	0.1354	1.0000	0.0000033

Bemerkung 13.15 *Um die Werte von* $\min S_0$ *und* $\max S_0$ *mit* σ_0^2 *zu vergleichen, müssen sie durch 0.9795 (im Fall* $c = 10$*) geteilt werden. Siehe dazu Bemerkung 12.2.*

Bemerkung 13.16 *Die KQ-FPCI, deren Existenz hier nachgewiesen wurde, würden auch erhalten bleiben, wenn neben* Q_1 *und* Q_2 *noch Mischungskomponenten vorhanden wären, die keine Masse auf die Menge* I_0 *legen (zum Beispiel einpunktverteilte Ausreißer). Das läßt sich analog zu den Sätzen in Abschnitt 13.1 aus Korollar 7.6 folgern. Entsprechendes gilt für Beispiel 13.6.*

Es zeigt sich, daß im Beispiel σ_0^2 extrem klein sein muß, um die Voraussetzungen des Satzes 13.11 zu erfüllen. Würde man zum Beispiel n Punkte nach Q_1 erzeugen, so wäre optisch nicht festzustellen, daß sie überhaupt von der durch β_1 definierten Gerade abweichen. Die Aussage ist insofern mehr als Grenzwertaussage interessant: Für $\sigma_0^2 \searrow 0$ existieren KQ-FPCI g_i, $i = 1, 2$ mit $\beta(g_i, Q) \in M_i$, $M_i \to \{\beta_i\}$. Solche Aussagen sind auf diese Weise auch für beliebige andere Mischungen von Normalverteilungen (Modell 3) zu erhalten, d.h. für andere Proportionen als $\frac{1}{2}$, andere β_i, die allerdings paarweise verschieden sein müssen, und mit etwas mehr Aufwand auch für mehr als zwei Komponenten und $p > 1$. In Analogie zum Lokationsfall folgen nun wieder Bedingungen für approximative Fisher-Konsistenz von $\beta(g, Q)$ für β_0. Dabei gelten die Bezeichnungen von Satz 13.11 und

$$Q(\epsilon, \sigma_0^2) := (1 - \epsilon)P + \epsilon H^*$$

mit P gemäß (13.27). Für gegebenes $t \in \left(0, \frac{1}{2}\right]$ sei immer

$$\epsilon^* = \epsilon^*(t) := \frac{0.04256t^2(1 + 0.8587t^2)}{4c(1 + t^2)(1 + 0.5025t^2) + (c - 1)0.04256t^2(1 + 0.8587t^2)}$$

so groß wie möglich. Falls t, σ_0^2 variabel sind, schreibe ich im folgenden auch $M_0(t)$ bzw. $M_0(t, \sigma_0^2)$ statt M_0, entsprechend für S_0 und I_0. c sei fest, so daß (13.33) erfüllt ist.

Hilfssatz 13.17 *Es gilt*

$$t \searrow 0 \Rightarrow M_0(t) \searrow \{\beta_0\}, \quad I_0(t) \searrow \{(y - x'\beta_0)^2 \leq \sigma_0^2\}, \quad \epsilon^*(t) \searrow 0,$$
$$\sigma_0 \searrow 0 \Rightarrow M_0(t, \sigma_0^2) \searrow \{\beta_0\}, \quad I_0(t, \sigma_0^2) \searrow \{(y - x'\beta_0)^2 = 0\},$$
$$t \to 0 \Rightarrow S_0(\epsilon^*) \to \left[\tfrac{\sigma_0^2}{c}, \sigma_0^2\right].$$

Beweis: Der Nenner $N(t)$ von $\epsilon^*(t)$ ist größer ist als $4c$ und für den Zähler $Z(t)$ gilt

$$Z'(t) = 2 * 0.04256t + 4 * 0.8587 * 0.04256t^3 > 4Z(t),$$

da $t \leq \frac{1}{2}$. Also mit $N'(t) < 8c(1.5025 + t^2)t + (c-1)t$:

$$\epsilon^{*\prime}(t) = \frac{Z'(t)N(t) - N'(t)Z(t)}{N(t)^2} > \frac{[16c - N'(t)]Z(t)}{N(t)^2} > 0.$$

Außerdem gilt $\lim_{t\to 0} N(t) = 4c$, $\lim_{t\to 0} Z(t) = 0$, also $t \searrow 0 \Rightarrow \epsilon^*(t) \searrow 0$. Damit folgen die Konvergenzaussagen über M_0 und S_0 nach Definition. Weiter steigt $\sup S_0(t, \sigma_0^2)$ streng monoton in t und σ_0^2 und fällt für $\sigma_0^2 \searrow 0$ gegen 0. Daraus folgen die Aussagen über I_0.

Korollar 13.18 *Existiert ein $0 < t \leq \frac{1}{2}$, so daß für H^* (13.36) und (13.37) für $(\theta, s^2) \in M_0(t) \times \left(\frac{\sigma_0^2}{c}, \sup S_0(t)\right]$ sowie (13.38) erfüllt sind, dann*

$$\exists \epsilon_1 > 0 \quad \forall \epsilon \leq \epsilon_1 \quad \exists t(\epsilon), \; KQ\text{-}FPCI\; g \;bzgl.\; Q(\epsilon, \sigma_0^2):$$
$$\beta[g, Q(\epsilon, \sigma_0^2)] \in M_0[t(\epsilon)], \quad \sigma^2[g, Q(\epsilon, \sigma_0^2)] \in S_0[t(\epsilon)].$$
$$\epsilon \searrow 0 \Rightarrow M_0[t(\epsilon)] \searrow \{\beta_0\}, \quad S_0[t(\epsilon)] \to \left[\tfrac{\sigma_0^2}{c}, \sigma_0^2\right].$$

Beweis: Mit P gemäß (13.27) gilt

$$\inf_{\theta \in M_0(t), s^2 \in S_0(t)} P\{(y - x'\theta)^2 \leq cs^2\} > P\{(|y| + \|x\|t)^2 \leq 1\} =: P_t,$$

da $\sqrt{c}s \geq 1$. Mit

$$\frac{H^*\{(y - x'\theta)^2 \leq cs^2\}}{P\{(y - x'\theta)^2 \leq cs^2\}} \leq \frac{1}{P_t} = \epsilon_0$$

ist offenbar (13.35) erfüllt. Für hinreichend kleines $\epsilon > 0$ kann $t(\epsilon)$ wegen Hilfssatz 13.17 und der Stetigkeit von $\epsilon^*(t(\epsilon))$ so gewählt werden, daß $\epsilon^*(t(\epsilon)) = \frac{\epsilon_0 \epsilon}{1-\epsilon}$. $t(\epsilon)$ konvergiert mit ϵ gegen 0 und die Voraussetzungen (13.36), (13.37) und (13.38) sind mit Hilfssatz 13.17 auch für $t_2 < t_1$ erfüllt, sobald sie für t_1 erfüllt sind, denn $\inf S_0(t_2) > \frac{\sigma_0^2}{c}$. Also folgt alles aus Satz 13.11 und Hilfssatz 13.17.

Korollar 13.19 *Existiert für gegebenes σ_1^2 ein $0 < t \leq \frac{1}{2}$, so daß für H^* (13.35), (13.36) und (13.37) für $(\theta, s^2) \in M_0(t, \sigma_1^2) \times (0, \sup S_0(t, \sigma_1^2)]$ sowie (13.38) erfüllt sind, dann*

$$\forall \sigma_2^2 \leq \sigma_1^2 \; \exists \; KQ\text{-}FPCI\; g: \; \beta[g, Q(\epsilon, \sigma_2^2)] \in M_0(t, \sigma_2^2), \; \sigma^2[g, Q(\epsilon, \sigma_0^2)] \in S_0(t, \sigma_2^2).$$
$$\sigma_2^2 \searrow 0 \Rightarrow M_0(t, \sigma_2^2) \searrow \{\beta_0\}.$$

Beweis: Nach den Voraussetzungen ist für ϵ_0, σ_1^2 Satz 13.11 anwendbar. Für $\sigma_2^2 < \sigma_1^2$ ist $S_0(t, \sigma_2^2) \subset (0, \max S_0(t, \sigma_1^2)]$ und mit Hilfssatz 13.17 damit auch $I_0(t, \sigma_2^2) \subset I_0(t, \sigma_1^2)$. Die Voraussetzungen für Satz 13.11 sind für σ_2^2 also für dieselben ϵ_0, t erfüllt und $M_0(t, \sigma_2^2) \searrow \{\beta_0\}$ mit $\sigma_2^2 \searrow 0$ folgt aus Hilfssatz 13.17.

Bemerkung 13.20 *Falls H^* auf einer Menge $\{(y - x'\beta_0)^2 \leq b\}$ mit $b > 0$ eine beschränkte λ^{p+1}–Dichte besitzt, sind die Voraussetzungen (13.35) und (13.36) für $(\theta, s^2) \in M_0(t, \sigma_1^2) \times (0, \sup S_0(t, \sigma_1^2)]$ für geeignete σ_1^2, t analog zu Korollar 13.9 erfüllt. Für die Anwendbarkeit von Korollar 13.19 müssen dann allerdings noch (13.37) und (13.38) gefordert werden.*

Teil IV
Simulationen

14 Einführung: Simulationen

14.1 Die Rolle der Simulationen bei der Beurteilung der Verfahren

Zur Analyse der hier interessierenden Datensätze bei unbekannter Clusterzahl scheinen mir im wesentlichen die drei Verfahren brauchbar zu ein, die schon in Abschnitt 10 verwendet wurden: Mischmodell-Maximum Likelihood-Schätzung (MML) gemäß Abschnitt 3.3, Fixed Partition Model-ML-Schätzung (FPML) gemäß Abschnitt 3.4, jeweils mit der dort diskutierten Schätzung der Clusterzahl, sowie die Fixpunktclusteranalyse (FPCA). Ich fasse kurz die vorhandenen theoretischen Resultate über die drei Verfahren zusammen:

- Für die FPCA steht ein konvergenter Algorithmus zur Verfügung, mit dem einige, aber nicht notwendig alle KQ-FPCV eines Datensatzes gefunden werden können. Ein KQ-FPCV ist ein kanonischer Schätzer eines KQ-FPCI. Die Existenz von KQ-FPCI bzgl. verschiedener Verteilungen wurde nachgewiesen. Die Abweichung ihrer Parameter $\beta(g,P), \sigma^2(g,P)$ von den entsprechenden Modellparametern einer Verteilungskomponente der Form (13.27) wurde beschränkt. Die Sätze sind aber nur anwendbar, wenn diese Verteilungskomponente sehr gut von den weiteren Verteilungsanteilen getrennt ist. Insbesondere wurde Existenz und Eindeutigkeit im homogenen Regressionsmodell bewiesen.

- Für MML steht ein Algorithmus zur Verfügung, der ein lokales Maximum der Likelihood liefert. Der Verdacht, das Verfahren liefere bei bekannter Clusterzahl konsistente und asymptotisch normale Parameterschätzer, ist begründet, aber nicht bewiesen. Bewiesen ist eine solche Aussage nur für den Lokationsfall. Ebenso stehen für die Schätzung der Anzahl der Cluster nur Resultate im Lokations- und Wechselpunktproblem zur Verfügung; nur im Wechselpunktproblem, das mit dem MML-Verfahren wenig zu tun hat, gibt es ein Konsistenzresultat.

- Für FPML steht ein Algorithmus zur Verfügung, der ein lokales Maximum der Likelihood liefert. Über das asymptotische Verhalten des Schätzers gibt es keine Ergebnisse. Er ist wie im Lokationsfall verzerrt. Die Verzerrung ließe sich unter Umständen abschätzen, ähnlich wie in Teil III dieser Arbeit für die FPCA. Für die Schätzung der Zahl der Cluster gibt es keine theoretische Untermauerung, auch nicht im Lokationsfall.

Die Theorie liefert also bislang wenig Anhaltspunkte für die Qualitätsbeurteilung der Verfahren. Um sich eine begründete Vorstellung davon zu machen, wie die Verfahren sich bei Daten aus den untersuchten Modellen verhalten, reichen die Ergebnisse nicht aus. Eine solche Vorstellung ist aber vonnöten, um die Verfahren in der Praxis sinnvoll einsetzen zu können, und um sich anhand vorgegebener Situationen für eine bestimmte

Methode zu entscheiden. Die Entfernung zwischen den theoretischen Ergebnissen und der Anwendung ist bei der Problemstellung dieser Arbeit also groß. Die Simulationen sollen helfen, diese Entfernung zu überbrücken.

Für die Anwendbarkeit der Verfahren sind folgende Fragen relevant, zu deren Beantwortung die Simulationen beitragen sollen:

- Wovon hängt es ab, ob ein Verfahren brauchbar ist oder nicht? Was sind die jeweiligen Stärken und Schwächen?
- Produzieren die Verfahren sinnvolle Resultate wenigstens in den idealen Modellsituationen, für die sie speziell entwickelt wurden?
- Wie sieht der Vergleich zwischen den Verfahren aus? Was ist in welcher Situation vorzuziehen?
- Was passiert, wenn man die Verfahren in Situationen einsetzt, für die sie nicht geschaffen sind? Konkret zum Beispiel: Kann man mit dem ML-Schätzer aus dem Mischmodell 1 auch sinnvoll Parameter schätzen, wenn die Zuordnungsunabhängigkeit verletzt ist (vgl. Bemerkung 2.2)? Was passiert bei Ausreißern?

Die Aussagefähigkeit von Simulationen für reale Datensituationen ist immer problematisch, da simulierte Datensätze in unrealistischer Weise den Modellvoraussetzungen folgen.

Der Zusammenhang zwischen Simulationen und Theorie ist ebenfalls problematisch, da mit Simulationen nur ausgewählte Spezialfälle untersucht werden können. Andererseits können die Simulationen Anhaltspunkte dafür geben, ob sich die vorhandenen theoretischen Resultate im Verhalten der Verfahren bei konkreten Datensätzen widerspiegeln und welche weiteren theoretischen Aussagen möglich erscheinen:

- Existieren in Datensätzen normalerweise FPCV, die den FPCI aus Teil III entsprechen?
- Was passiert bei steigender Stichprobengröße? Machen die MML-Parameterschätzungen und die Clusterzahlschätzungen einen konsistenten Eindruck?
- Wie wirkt sich die Verzerrung der FPML-Schätzer und FPCA-Parameter bei überlappenden Clustern aus? Verschwindet sie im Vergleich zur Varianz?

Für das MML-Verfahren führten DeSarbo und Cron (1989) eine Simulation durch. Diese Simulation ist allerdings nicht mit meinen Simulationen vergleichbar, da die Autoren von bekannter Clusterzahl ausgingen und andere Kenngrößen verwendeten (siehe Abschnitt 14.2).

14.2 Überlegungen zum Versuchsaufbau

In Abschnitt 15 werden die Simulationen genau beschrieben, d.h. die verglichenen Verfahren, die Modellsituationen und die erhobenen Statistiken. Hierzu nun einige Vorüberlegungen:

Alle Verfahren werden jeweils nur in einer Variante verglichen. Die Frage nach optimaler Wahl der für ein bestimmtes Verfahren verwendeten Parameter (Anzahl der

Iterationen pro Datensatz; Wahl von c bei den Fixpunktclustern) wird nicht behandelt. Weiterhin beschränke ich mich auf den Fall unbekannter Clusterzahl; anderenfalls würde die Anwendung von Fixpunktclustern keinen Sinn ergeben, denn wesentliche Information bliebe ungenutzt. Die Störterme werden immer unabhängig normalverteilt gewählt.

Die Anzahl der denkbaren interessanten Modellsituationen für die Simulationen ist sehr hoch. Die Dimension p, die Stichprobengröße n und die Anzahl der Komponenten s sind freie Parameter. Die Modellcluster[20] können gleiche oder unterschiedliche Größe und gleiche oder unterschiedliche Störvarianz σ^2 haben. Weiterhin hängen die Ergebnisse stark von den Regressionsparametern ab. Diese Parameter bestimmen, wie die Cluster sich zueinander verhalten: Sie können zum Beispiel parallel liegen oder sich kreuzen und dann entsprechend größere Überschneidungen haben. Auch die Frage, ob und wie stark sich die Regressoren für die einzelnen Cluster unterscheiden, spielt eine große Rolle.

Bei den hier durchgeführten Simulationen liegt der Schwerpunkt auf Situationen mit gleichartigen Clustern (gleiche Größe, gleiche Störvarianz). Die Regressionsparameter werden meistens zufällig gewählt, spezielle Situationen mit zwei Clustern (parallel - über Kreuz) werden aber mit vorgegebenen Regressionsparametern simuliert, um den Einfluß der Konstellation zu untersuchen. Fast alle Situationen werden mit vier verschiedenen Stichprobengrößen simuliert, so daß die Entwicklung der Ergebnisse bei steigendem n zu verfolgen ist.

Fast alle Simulationen beruhen auf Daten, die nach Modell 4 erzeugt werden, d.h. es gibt immer eine feste Partition der Punkte zu den Clustern und die Regressoren sind zufällig. Zwar wird dieses Modell in der Theorie nicht behandelt, aber es eignet sich gut für allgemeine Vergleiche. Zum einen muß keine bestimmte Konstellation von Regressoren vorausgesetzt werden[21]. Zum anderen ermöglicht eine feste Partition eine genaue Untersuchung der Zuordnung der Punkte zu den Clustern. Das ist mit den Clusterproportionen aus Mischmodellen nicht so einfach möglich. Die Klassifikation der Punkte ist interessanter als die Proportionsschätzung[22], weil letztere auch zufällig korrekt sein kann, wenn die sonstigen Parameterschätzungen eine große Abweichung von den Modellparametern haben und also eigentlich ganz andere Cluster geschätzt werden. Die Proportionsschätzung wird in den Simulationen nicht untersucht.

Vor allem interessieren mich Modellsituationen, in denen ein deutliches Muster zu finden ist. Im Hinblick auf die Brauchbarkeit der Verfahren zur Analyse von Daten ist für mich die Erkennung deutlicher Muster ein wichtigeres Kriterium als die relativ gute Parameterschätzung in Situationen, in denen fragwürdig ist, ob mehrere lineare Regressionscluster ein angemessenes Modell zur Beschreibung des Datensatzes liefern. Daher wird die Störvarianz meistens im Vergleich zur Varianz der Regressoren sehr klein gewählt (0.01 oder 0.001 zu 1). Zur Illustration: Abbildung 10 zeigt links ein Beispiel mit drei gleich großen Clustern, jeder mit Störvarianz 0.1. Der Datensatz wurde nach Konstellation 1randl3[23] erzeugt. Die Daten rechts entstanden im Prinzip nach demselben Mechanismus (1randl13), nur mit Störvarianz 0.01. Man kann sich darüber streiten,

[20] Die Regressionsverteilungen, die in einer Fixed Partition-Verteilung - hier entsprechend Modell 4 - als Randverteilungen für die einzelnen $i \in I$ auftauchen, definieren die „Modellcluster". Ein Modellcluster ist bestimmt durch $\beta, \sigma^2, G, \{i \in I : \gamma(i) = (\beta, \sigma^2, G)\}$.

[21] Allerdings ist eine Verteilungsvoraussetzung natürlich auch willkürlich.

[22] Die Proportionsschätzung ist die Schätzung der Mischungsanteile $\epsilon_i, i = 1, \ldots, s$ in den Mischmodellen 3 und 1.

[23] Näheres zur Schreibweise findet sich in der Beschreibung der Konstellationen in Abschnitt 15.2.

ob im linken Fall ein „deutliches Muster" vorliegt. Mehrere unabhängige menschliche Testaugen rieten angesichts dieses Bildes die richtige Konstellation. Andererseits scheint das menschliche Auge auch dazu zu neigen, implizit und für diesen Fall korrekterweise gleiche Störvarianzen vorauszusetzen, was die Verfahren nicht wissen. Jedenfalls ist es offenbar nicht nur für ein automatisiertes Verfahren schwierig, eine vernünftige Klassifikation der Punkte in die Clustern vorzunehmen. Meistens wird in den Simulationen also mit kleineren Störvarianzen, d.h. deutlicheren Mustern gearbeitet. In Abschnitt 16.3 findet man jedoch auch die Ergebnisse der Simulation mit 1randl3.

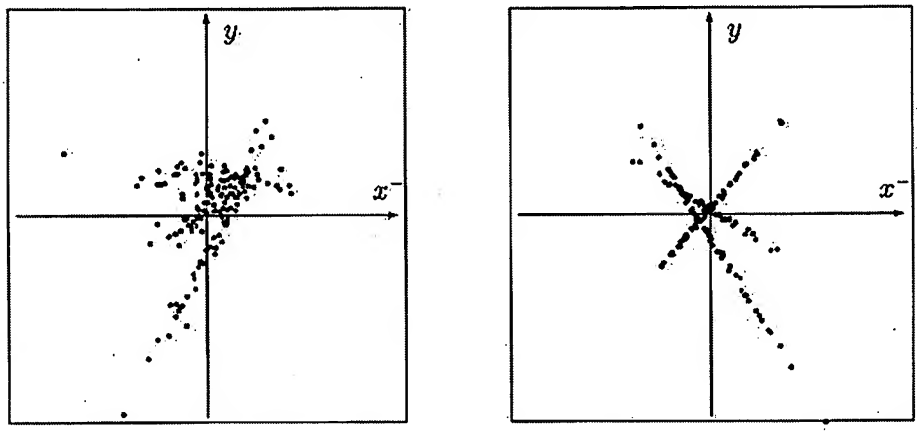


Abbildung 10: Daten aus 1randl3 und 1randll3 ($n(1) = 50$)

Modell 4 hängt mit den anderen Modellen folgendermaßen zusammen: Bedingt unter festem $(x_i)_{i\in I}$ ist $(y_i)_{i\in I}$ gemäß Modell 2 verteilt, also nach den Voraussetzungen der Fixed Partition-ML-Schätzung. Wären die Clusterzuordnungen $\gamma(i)$, $i \in I$ unabhängig multinomialverteilt, so wäre (x_i, y_i) nach Modell 3 verteilt. Wären die Regressorenverteilungen G dann noch für alle $(\beta, \sigma^2, G) \in \gamma(I)$ gleich, so hätte $(y_i)_{i\in I}$ bedingt unter gegebenem $(x_i)_{i\in I}$ eine Verteilung gemäß Modell 1, also entsprechend den Voraussetzungen der Mischmodell-ML-Schätzung.

Zu den erhobenen Statistiken: Ich habe mich gegen Verzerrung (Bias) und mittleren quadratischen Fehler (MSE) der Parameterschätzungen als Kenngrößen entschieden, denn diese Maße sind höchst unrobust. Wenn ein Verfahren in einer bestimmten Clustersituation sehr selten die Konstellation der Cluster völlig falsch einschätzt, beeinflussen diese Fälle den Bias und MSE der ganzen Simulation. Letztlich würde dann nicht gemessen, wie häufig ein Verfahren gut abschneidet, sondern nur, wie schlecht es abschneidet, wenn es schlecht abschneidet. Dieser Effekt erzeugt zum Beispiel in den Simulationen von DeSarbo und Cron (1988) häufig sehr hohe MSEs. Weiter ist zu berücksichtigen, daß hier Verfahren verglichen werden, die keine einheitliche Ausgabe haben: FPML schätzt eine Partition, MML schätzt Clusterproportionen und die von Fixpunktclustern geschätzte Clusterzahl ist von anderer Qualität als die der anderen Verfahren, weil beliebige Überschneidungen erlaubt sind. Zum Beispiel könnte das Fixpunktclusterverfahren in einer Konstellation mit drei Clustern neben den drei korrekten Clustern noch sieben weitere finden.

Ich arbeite daher mit „gröberen" Statistiken, die beschreiben sollen, ob das Verfahren die allgemeine Struktur des Datensatzes erkennt. Es wird nur festgestellt, ob ein

Verfahren einen bestimmten Cluster korrekt findet oder nicht. Dafür werden dann die Häufigkeiten erhoben. Was heißt nun „korrekt gefunden"? Dafür werden mehrere Kriterien verwendet, die in Abschnitt 15.3 formal definiert werden:

- Das „β-Kriterium" entscheidet danach, ob der geschätzte Regressionsparameter $\hat{\beta}$ den entsprechenden Clusterparameter gut approximiert. Es ist sinnvoll, ein Kriterium zu verwenden, das nicht von $\hat{\sigma}^2$ abhängt, denn es ist möglich, daß die Überschneidung der Cluster den Störvarianzschätzer stark verzerrt, auch wenn die grobe Clusterstruktur richtig erkannt wird.

- Das „$\beta - \sigma$-Kriterium" entscheidet danach, ob Regressions- und Skalenparameter gut approximiert werden.

- Das „Zuordnungskriterium" hängt von der Zahl der fehlklassifizierten Punkte ab. Das sind die Punkte, die nach Modellvoraussetzung zum Cluster gehören müßten, aber nicht hereinklassifiziert werden und die Punkte, die dem Cluster zugeordnet werden, aber nach Modellvoraussetzung nicht dazugehören. Während die ersten beiden Kriterien die Qualität der Parameterschätzungen messen, mißt das Zuordnungskriterium die Qualität der Klassifikation. Je nach Anwendung kann das eine oder das andere vorrangiges Ziel der Datenanalyse sein.

Alle vom Verfahren gefundenen Cluster werden mit allen Modellclustern verglichen. Das bedeutet insbesondere, daß die ML-Verfahren Cluster korrekt finden können, wenn sie die Clusterzahl falsch einschätzen. Auch kann theoretisch, bei sehr ähnlichen Parametern der Modellcluster, ein einzelner vom Verfahren gefundener Cluster mehrere Modellcluster gleichzeitig korrekt finden.

Außerdem wird die Verteilung der Anzahl der gefundenen Cluster bzw. die geschätzte Clusterzahl erhoben. Zu beachten ist, daß die Vergleichbarkeit der Statistiken über Fixpunktcluster begrenzt ist: Weil das Verfahren normalerweise mehr Cluster findet als die „korrekten", was durchaus dem Sinn der FPCA entspricht, darf die Anzahl der gefundenen Cluster nicht als Schätzung interpretiert werden. Es ist zu erwarten, daß sie höher liegt als bei den ML-Verfahren. Wenn die FPCA einen Cluster korrekt findet, kann es andererseits schwierig sein, ihn in der Ausgabe des Verfahrens als „relevant" zu erkennen, weil er unter Umständen unter vielen weiteren gefundenen Fixpunktclustern versteckt ist. Die Simulationsergebnisse für die Parameter der korrekt gefundenen Cluster sind also bei Fixpunktclustern kritischer zu beurteilen als bei den anderen Verfahren, die Clusteranzahlen hingegen großzügiger.

Nicht jede der erhobenen Statistiken ist in jedem Fall sinnvoll. Zum Beispiel sind in vielen Datenkonstellationen alle Cluster symmetrisch, d.h. sie haben dieselbe Größe und Störvarianz sowie symmetrische Regressionsparameter. In diesen Fällen ist es sinnlos, die Findungshäufigkeiten nach den einzelnen Clustern aufzuschlüsseln, weil die Situation für alle Cluster symmetrisch ist.

Abschnitt 16 enthält nicht die vollständigen Ausgaben der Simulationen, sondern die Ergebnisse sind komprimiert. Die Zusammenfassungen werden dort aber jeweils explizit erklärt.

15 Beschreibung der Simulationen

15.1 Die Verfahren

15.1.1 Fixpunktclusteranalyse (FPCA)

KQ-Fixpunktclustervektoren aus Abschnitt 8.2 werden nach Algorithmus 2 aus Abschnitt 9 berechnet. Der Algorithmus konvergiert für jeden der in der Simulation generierten Datensätze. Zuerst wird die Iteration mit $g^0 \equiv 1$ gestartet, also mit dem kompletten Datensatz. Danach werden für jede weitere Iteration $p+3$ Punkte $(x_{i_k}, y_{i_k})_{k\in\{1,\ldots,p+3\}}$ zufällig ausgewählt, so daß $g^0_{i_k} = 1$ falls $k \in \{1,\ldots,p+3\}$ und $g^0_i = 0$ sonst. Die Anzahl der Iterationen pro Datensatz beträgt $100 + 40p^2$.

Bemerkung 15.1 *Angenommen ein Datensatz hätte Stichprobengröße n und enthielte einen Modellcluster der Größe $n_1 < n$. Weiter angenommen, es gäbe einen FPC, der diesem Modellcluster approximativ entspräche. Meine Erfahrung zeigt, daß der Algorithmus fast immer diesen FPC findet, wenn alle m Punkte des Iterationsstartes ($m < n$) aus den n_1 Punkten des Modellclusters kommen, und fast nie sonst. Die Wahrscheinlichkeit, beim Ziehen ohne Zurücklegen die m Punkte genau aus den n_1 Punkten des Clusters zu wählen, ist*

$$P(m) := \frac{\binom{n_1}{m}}{\binom{n}{m}} = \frac{n_1!(n-m)!}{n!(n_1-m)!}.$$

Ist n_1 nur wenig größer als m, so ist $P(m)$ extrem klein. Deshalb ist es sinnvoll, die Anzahl der Punkte, mit denen die Iteration beginnt, so klein wie möglich zu wählen. Ich wähle $m = p+3$, weil bei $m = p+2$ die Wahrscheinlichkeit noch sehr groß ist, für den ersten KQ-Schätzer, der nur auf diesen m Punkten beruht, annähernd Residuenskala 0 zu erhalten.

Was passiert nun für steigendes p, zum Beispiel beim Übergang von p auf $p+1$? Es ist $\frac{P((p+1)+3)}{P(p+3)} = \frac{n_1-p-3}{n-p-3} \approx \frac{n_1}{n}$, wenn n und n_1 gemessen an p groß sind. Die Wahrscheinlichkeit, einen Cluster der Größe n_1 aus n Punkten zu finden, wird also bei steigendem p exponentiell kleiner.

Wollte man zum Beispiel die Häufigkeit, mit der ein Cluster mit $\frac{n_1}{n} = \frac{1}{3}$ in großen Stichproben im Mittel gefunden wird, approximativ gleichhalten, so würde sich $k3^p$ mit geeignetem k als Iterationsanzahl anbieten. Das führt aber bei großem p zu nicht mehr akzeptablen Rechenzeiten (siehe Abschnitt 15.1.4). Ich habe mich daher für quadratisches Wachstum entschieden.

Als Mindestclustergröße wird $2(p+3)$ gewählt; Iterationsdurchgänge, die auf kleinere Cluster führen, werden in der Ausgabe ignoriert. Die Festlegung einer Mindestclustergröße ist sinnvoll, da sehr kleine Teildatensätze häufig zufällig sehr gut durch eine gemeinsame Hyperebene angepaßt werden können und dadurch einen irrelevanten FPC bilden. Weiter ist $c = 10$ (siehe dazu Abschnitt 8.2).

Die Skalenschätzung $\sigma^2(g)$ für jeden FPCV g wurde entsprechend Bemerkung 12.2 durch 0.9795 geteilt, um für Fisherkonsistenz im Falle homogener Populationen zu sorgen.

15.1.2 Mischmodell-ML (MML)

Die Lösungen des Mischmodell-ML-Verfahrens werden mit dem EM-Algorithmus nach DeSarbo und Cron (1988) aus Abschnitt 3.3 berechnet. Dabei werden nacheinander die Lösungen für 1 bis 6 Cluster berechnet, bzw. bis 4 Cluster falls $n < 6(p+2)$. Die optimale Clusterzahl wird mit dem BIC (3.1) berechnet.

Bemerkung 15.2 *Die Schätzung der Clusterzahl mit dem BIC führt in 94 Fällen zu einer zum Teil deutlichen Über- und in 7 Fällen zu einer leichten Unterschätzung der Clusterzahl, jeweils im Mittel über die Simulationsläufe. Neunmal wurde die richtige Clusterzahl im Mittel auf zwei Nachkommastellen genau getroffen. Nach Bemerkung 3.2 ist daher zu erwarten, daß das AIC (3.4) noch deutlich schlechter abgeschnitten hätte. Das hat sich in einigen Testdurchläufen auch bestätigt; insbesondere scheint das AIC unbrauchbar zu sein, um eine homogene Population - nur ein Cluster - zu erkennen, falls* $n \leq 100$.

Aus Zeitgründen wird pro Datensatz und Clusterzahl nur eine Iteration durchgeführt (siehe Abschnitt 15.1.4). Bei mehreren Iterationen wären bessere Ergebnisse möglich gewesen, denn die Wahrscheinlichkeit, das globale Maximum der Likelihood zu finden, wäre höher gewesen. Zur Startpartition: Für $i = 1, \ldots, n$ wird gleichverteilt $j \in \{1, \ldots, s\}$ ausgelost, so daß $\hat{\epsilon}_{ij} = 1$. Die anderen $\hat{\epsilon}_{kl}$ werden gleich Null gesetzt. Dieses Verfahren wird so lange wiederholt, bis jedem der s Cluster mindestens $p + 2$ Punkte zugeteilt sind. Als untere Schranke für die $\hat{\sigma}_j^2$, $j \in \{1, \ldots, s\}$ wird 10^{-3} gewählt. Eine Iteration wird beendet, wenn sich die Loglikelihood in einem Iterationsschritt um weniger als 10^{-6} verbessert. Um numerische Probleme zu vermeiden, wird die Iteration abgebrochen und nicht wiederholt, wenn der Anteil $\hat{\epsilon}_j$, $j \in \{1, \ldots, s\}$ für einen Cluster kleiner als 10^{-4} ist oder es für einen Cluster $j \in \{1, \ldots, s\}$ weniger als $p + 2$ Punkte gibt, so daß $\hat{\epsilon}_{ij} \geq 10^{-5}$, $i \in \{1, \ldots, n\}$. In diesen Fällen wird die entsprechende Clusterzahl als nicht optimal gewertet.

Für die Statistiken in der Simulation wird für das Mischmodell-ML-Verfahren eine Klassifikation der Punkte in die Clustern benötigt, die nach beendeter Iteration für $i = 1, \ldots, n$ vorgenommen wird gemäß

$$\hat{\zeta}_{MML}(i) := \arg\max_{j=1,\ldots,s} \hat{\epsilon}_{ij}. \tag{15.1}$$

15.1.3 Fixed Partition-ML (FPML)

Die Lösungen des Fixed Partition-ML-Verfahrens werden mit dem in Abschnitt 3.4 beschriebenen Algorithmus berechnet. Dabei werden nacheinander die Lösungen für 1 bis 6 Cluster berechnet, bzw. bis 4 Cluster falls $n < 6(p + 2)$. Die optimale Clusterzahl wird mit dem modifizierten BIC (3.11) berechnet. Die Anzahl der Iterationsdurchgänge pro Datensatz und Clusterzahl ist abgerundet $\frac{s(4+p)}{1+n/100}$, wobei s die aktuelle Clusterzahl bezeichnet.

Bemerkung 15.3 *Eine testweise Vergrößerung der Anzahl der Iterationsdurchgänge bringt keine nennenswerten Verbesserungen; aus Zeitgründen wäre das aber unproblematisch gewesen (siehe Abschnitt 15.1.4). Allgemein benötigt das Verfahren mit steigendem*

s mehr und mit steigendem n weniger Iterationen, um zu guten Lösungen zu kommen. Das spiegelt sich in der von mir gewählten Zahl der Durchgänge wieder.

Die Ergebnisse der Schätzung der Anzahl der Cluster durch das Kriterium (3.11) sind ausgezeichnet. Ich habe vorher das AIC und BIC getestet, wobei ich $k(s)$ wie im Mischmodell-ML-Verfahren (Abschnitt 3.3) gewählt habe. Die Ergebnisse waren unbrauchbar: Die Clusteranzahl wurde fast immer maximal möglich (d.h. hier 6) geschätzt. Das liegt daran, daß die Parameter $\zeta(i)$, $i = 1, \ldots, n$ von einer anderen Art sind als die anderen reellen Parameter: Ihre Anzahl wächst mit n, sie sind ganzzahlig und ihr Wertebreich wächst mit s. Mit Kriterium (3.11) habe ich versucht, das zu berücksichtigen. Ich habe auch einige Tests durchgeführt, in denen statt des Faktors 0.7 die Werte 0.5 und 1 verwendet wurden, was eher zu Über- bzw. Unterschätzungen führte. Es ist aber festzuhalten, daß ich die Wahl des Kriteriums (3.11) für kaum begründet halte. Es funktioniert nur überraschend gut. Dabei ist zu beachten, daß möglicherweise eine Erhöhung der Iterationszahl dazu führt, daß die Clusterzahl vom Verfahren fälschlicherweise höher eingeschätzt wird: Die Wahrscheinlichkeit, zufällig Lösungen zu finden, in denen einzelne kleine Cluster eine sehr kleine Residuenskala haben, steigt bei mehr Iterationen und mehr Clustern. Daraus würde sich eine sehr hohe Likelihood und damit eventuell eine Überschätzung der Clusterzahl ergeben. Die nur empirisch gezeigte Qualität des Kriteriums (3.11) hängt also möglicherweise von der Iterationszahl ab.

Die Startpartition wird so ermittelt wie im Mischmodell-ML-Verfahren, wobei $\hat{\zeta}(i) = j :\Leftrightarrow \hat{c}_{ij} = 1$. Die Iteration wird abgebrochen und nicht wiederholt, wenn Cluster auftauchen, die weniger als $p + 2$ Punkte enthalten. Falls alle Iterationen für eine bestimmte Clusterzahl mit diesem Ereignis enden, wird die Clusterzahl als nicht optimal gewertet. Ansonsten gibt es keine numerischen Probleme.

15.1.4 Geschwindigkeitsvergleich

Um einen Eindruck zu bekommen, wie schnell sich die Verfahren rechnen lassen, habe ich in vier Simulations-Situationen die Rechenzeiten (CPU-Zeit auf IBM Risc/6000 PowerPC 250 in Minuten:Sekunden) nachgemessen. In diesen Situationen sind die Cluster gleich groß, die Regressoren sind verteilt nach $\mathcal{N}_{(0,1)}$, ebenso die Regressionsparameter. Die Störvarianz ist 0.1 für alle Cluster. Die Verfahren werden wie oben beschrieben durchgeführt. Das heißt auch, daß für die ML-Verfahren alle Clusterzahlen von 1 bis 6 durchgerechnet werden. Die Zeiten verstehen sich inklusive der Generierung der Datensätze.

p	1	1	4	9
Clusterzahl	2	3	1	2
n	40	600	100	200
Anzahl simulierter Datensätze	50	10	10	10
FPCA	0:16	1:40	1:18	37:09
MML	0:28	22:23	0:28	1:14
FPML	0:12	0:51	0:14	1:06

Der FPML-Algorithmus ist durchgehend am schnellsten, obwohl im Unterschied zum MML-Algorithmus jeweils mehrere Iterationsdurchgänge durchgeführt wurden. Letzte-

rer konvergiert sehr langsam und jeder Durchgang erfordert für jeden Cluster die Berechnung einer gewichteten Regression mit dem kompletten Datensatz. Daher fällt hier besonders die Stichprobengröße n ins Gewicht. Das Ergebnis für $p = 9$ für die FPCA ist auf die oben bereits diskutierte Explosion der Zahl der Iterationen zurückzuführen. Auch in der Situation mit $p = 9$ benötigt eine einzelne Iteration nur knapp 0.07 Sekunden. Die Ergebnisse machen sehr deutlich, warum eine nennenswerte Erhöhung der Iterationsanzahlen bei MML und FPCA nicht mehr machbar gewesen wäre.

15.2 Die Erzeugung der Testdaten

Es gelten die Bezeichnungen aus Modell 4. In jeder Simulation werden die Datensätze nach einer bestimmten „Konstellation" erzeugt, die durch folgende Charakteristika festgelegt ist:

- Die Dimension p;
- die Anzahl der Cluster $s = |\gamma(I)|$ mit $\gamma(I) = \{(\beta_1, \sigma_1^2, G_1), \ldots, (\beta_s, \sigma_s^2, G_s)\}$;
- die Anzahlen der Punkte der einzelnen Cluster, im folgenden mit $n(1), \ldots, n(s)$ der Größe nach geordnet bezeichnet, wobei $n(1)$ die Größe des größten Clusters sei;
- die Regressorenverteilung G_j für jeden Cluster $j = 1, \ldots, s$;
- die Regressionsparameter β_j für jeden Cluster $j = 1, \ldots, s$;
- die Störvarianz σ_j^2 für jeden Cluster $j = 1, \ldots, s$.

Aufgrund der großen Zahl interessanter Möglichkeiten verwende ich keinen kombinatorisch vollständigen Versuchsplan. Das Ziel ist, viele verschiedene Konstellationen so auszuprobieren, daß ein guter Überblick über die Einflüsse der verschiedenen Charakteristika entsteht. Die Dimensionen $p = 1, 2, 4, 9$ werden einigermaßen gleichmäßig verwendet, die niedrigeren Dimensionen aber etwas häufiger wegen der höheren Anschaulichkeit der simulierten Konstellationen. Um einen guten Eindruck vom Einfluß der Stichprobengröße zu bekommen, wird in fast allen Fällen mit $n(1) = 20, 50, 100, 300$ simuliert. Es erscheint mir sinnvoll, $n = |I|$ nur indirekt über $n(1)$ zu variieren, d.h. es wird in der homogenen Situation mit bis zu 300, in der Situation mit drei gleichgroßen Clustern aber mit bis zu 900 Punkten gerechnet. Da die Verfahren Parameterschätzer innerhalb der einzelnen Cluster berechnen, hängt die Genauigkeit nicht von der Gesamtstichprobengröße ab, sondern von der Größe der Cluster. Entsprechend mehr Punkte bräuchte man auch in der Anwendung, um gleiche Präzision zu erreichen, wenn mehrere Cluster vorlägen.

Ich stelle nun die Datenkonstellationen vor, die simuliert werden. Die in Klammern angegebenen Worte im Schrifttyp sans serif geben die Bezeichnungen für die später erläuterte Kurzschreibweise an.

- Homogene Population ($s = 1$):
  - Normalverteilte Regressoren (hom): $G_1 = \mathcal{N}_{(0,I_p)}$, $\beta_1 = (0, 0)$, $\sigma_1^2 = 1$. Diese Konstellation wird simuliert für $p = 1, 2, 4, 9$, $n = n(1) = 20$ falls $p \neq 9$ sowie $n = 50, 100, 300$ in allen Fällen. Auf diese Situation läßt sich Satz 12.1 anwenden.

- Weitere Konstellationen mit festen Parameterwerten:
  - Zwei Lokationscluster mit Abstand 5 (lok): $s = 2$, $p = 0$, $\beta_1 = 0, \beta_2 = 5$, $\sigma_1^2 = \sigma_2^2 = 1$, $n(1) = n(2) = 20, 50, 100, 300$. Die Regressorenverteilung entfällt bei $p = 0$. Dieses ist die Situation, die in Beispiel 13.6 behandelt wird, allerdings mit fester Partition und einer kleineren Differenz zwischen den Clustermitteln, so daß es bereits eine sichtbare Überschneidung zwischen den Clustern gibt.
  - Kreuzförmige Konstellation (cross): $s = 2$, $p = 2$, $G_1 = G_2 = \mathcal{N}_{(0,I_p)}$, $\beta_1 = (1, 0, 0)$, $\beta_2 = (-1, 0, 0)$, $\sigma_1^2 = \sigma_2^2 = 0.01$, $n(1) = n(2) = 20, 50, 100, 300$. Beispiel 13.14 behandelt eine kreuzförmige Konstellation theoretisch, allerdings unter Voraussetzung einer winzigen, für die Anwendung irrelevanten Störvarianz. Im Unterschied dazu ist hier die Störvarianz größer und es kommt durch $p = 2$ noch die zusätzliche Schwierigkeit hinzu, daß die Verfahren die Nullen bei β_{i2}, β_{i3}, $i = 1, 2$, mitschätzen müssen.
  - Parallele Konstellation (par): $s = 2$, $p = 2$, $G_1 = G_2 = \mathcal{N}_{(0,I_p)}$, $\beta_1 = (0, 0, 0)$, $\beta_2 = (0, 0, 2)$, $\sigma_1^2 = \sigma_2^2 = 0.1$, $n(1) = n(2) = 20, 50, 100, 300$. $\|\beta_1 - \beta_2\| = 2$ gilt hier also wie bei cross. Die Störvarianzen sind größer, so daß die Ergebnisse nicht direkt vergleichbar sind. Mit $\sigma_1^2 = \sigma_2^2 = 0.01$ wäre die Konstellation aber so klar gewesen, daß alle Verfahren annähernd so gut gewesen wären wie die KQ-Schätzer für die einzelnen Cluster. Dadurch hätte man die Verfahren nicht mehr sinnvoll vergleichen können. Aufgrund der Äquivarianzeigenschaften der Verfahren gelten die hier erzielten Simulationsergebnisse auch, falls y für alle Daten mit $\sqrt{0.1} = 0.316$ multipliziert worden wäre, womit im Gesamtmodell $\sigma_1^2 = \sigma_2^2 = 0.01$ und $\beta_2 = (0, 0, 0.632)$ gegolten hätte (siehe Bemerkung 2.4).
  - Identifizierbarkeitsproblem (id): $s = 9$, $p = 1$, $G_1 = G_2 = G_3 = \mathcal{N}_{(0,0.001)}$, $G_4 = G_5 = G_6 = \mathcal{N}_{(1,0.001)}$, $G_7 = G_8 = G_9 = \mathcal{N}_{(2,0.001)}$, $\beta_1 = (0, 2)$, $\beta_2 = (0, 1)$, $\beta_3 = \beta_5 = (0.5, 0)$, $\beta_4 = \beta_7 = (1, 1)$, $\beta_6 = \beta_8 = (1, 0)$, $\beta_9 = (-0.5, 2)$, $\sigma_j^2 = 0.001$, $n(j) = 20, 50$, $j = 1, \ldots, 9$. Diese Konstellation ist aus Beispiel 5.5 abgeleitet. Dabei wird hier für jeden der neun Regressorenpunkte ein Cluster definiert, wobei die Datenpunkte in der Simulation sowohl in x- als auch in y-Richtung eine Varianz von 0.001 hatten. Versteht man unter einem Regressionscluster ein Modell mit gleichen Regressionsparametern und Störvarianzen, so gibt es hier sechs Cluster. Drei davon haben je zwei verschiedene Regressorenverteilungen[24]. Diese sechs Cluster teilen sich auf in die zweimal drei Cluster, deren Parameter in Beispiel 4.11 jeweils schon das komplette Modell beschreiben. Daher war zu erwarten, daß die ML-Verfahren häufig eine drei-Cluster-Lösung angeben. In der Diskussion der Ergebnisse unterscheide ich zwischen neun „Modellclustern" und sechs „echten Clustern".

  Für die Konstellationen lok, cross und par ist die Zuordnungsunabhängigkeit (Bemerkung 2.2) erfüllt, für id zumindest approximativ (bis auf die Regressorenvarianz 0.001).

- Gleichartige Cluster mit zufälligen Regressionsparametern:

[24] Bei dieser Ausdrucksweise wird sozusagen die Äquivalenzrelation „$\sim_s$" aus Beispiel 4.11 benutzt.

- Alle Regressorenverteilungen gleich (rand): $s = 2,3,4$, $G_j = \mathcal{N}_{(0,I_p)}$ für $j = 1,\ldots,s$. β_j wird für $j = 1,\ldots,s$ für jeden Simulationslauf unabhängig zufällig aus $\mathcal{N}_{(0,I_{p+1})}$ gewählt. Für $j = 1,\ldots,s$ sind die $\sigma_j^2, n(j)$ immer für alle Cluster gleich, $n(j) = 20,50,100,300$ außer für $p = 9$, dort $n(j) = 50,150$. Folgende Situationen für p, s, σ_j^2 werden simuliert:

s	2	2	3	3	3	4
p	1	4	1	4	9	1
σ_j^2	0.01	0.001	0.1	0.001	0.001	0.001

  In diesem Fall ist die für MML vorausgesetzte Zuordnungsunabhängigkeit erfüllt.

- Regressorenverteilungen unterschiedlich (randx): Wie Situation rand, außer $G_j = \mathcal{N}_{(a_j,I_p)}$ für $j = 1,\ldots,s$, wobei a_j für $j = 1,\ldots,s$ und für jeden Simulationslauf unabhängig zufällig aus $\mathcal{N}_{(0,\tau^2 I_p)}$ gewählt wird mit $\tau^2 = 9$. $n(j) = 20,50,100,300$ außer für $s = p = 4$, dort $n(j) = 20,50,150$. Folgende Situationen für p, s, σ_j^2 werden simuliert:

s	2	2	3	3	4
p	2	4	1	2	4
σ_j^2	0.01	0.001	0.01	0.001	0.001

- Regressorenverteilungen extrem unterschiedlich (rand!x): Wie Situation randx, nur mit $\tau^2 = 100$. $s = 2$, $p = 2$ mit $\sigma_j^2 = 0.01$, $n(1) = 100$ und $p = 9$ mit $\sigma^2 = 0.001$, $n(1) = 150$.

- Wechselpunkt-Situation (change): $s = 2$, $p = 1$, Regressoren für den ersten Cluster äquidistant und fest zwischen -1 und 0, Regressoren für den zweiten Cluster äquidistant und fest zwischen 0 und 1, $\sigma_1^2 = \sigma_2^2 = 0.01$, Regressionsparameter wie in Situation rand, $n(1) = n(2) = 20,50,100,300$.

- Verschiedenartige Cluster:

  - Zweiter Cluster halb so groß (halb): Regressoren und Regressionsparameter wie in Situation rand, $s = 2, p = 4$, $n(1) = 50,100,300$, $n(2) = \frac{n(1)}{2}$, $\sigma_1^2 = \sigma_2^2 = 0.001$.

  - Zwei Cluster mit unterschiedlichen Störskalen (sc): Regressoren und Regressionsparameter wie in Situation rand, $s = 2, p = 2, n(1) = n(2) = 20,50,100,300$, $\sigma_1^2 = 0.1$, $\sigma_2^2 = 0.001$.

  - Zweiter Cluster mit halber Größe. Störskalen und Regressoren sind unterschiedlich (schalbx): Regressoren und Regressionsparameter wie in Situation randx, $s = 2, p = 9$, $n(1) = 50,100,300$, $n(2) = \frac{n(1)}{2}$, $\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.001$.

  - Drei Cluster unterschiedlicher Größe. Störskalen und Regressoren sind unterschiedlich (scvx): Regressoren und Regressionsparameter wie in Situation randx, $s = 3, p = 1$, $n(1) = 20,50,100,300$, $n(2) = \frac{4n(1)}{5}$, $n(3) = \frac{3n(1)}{5}$, $\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.001$, $\sigma_3^2 = 0.003$.

- Ausreißerkonstellationen: Im Rahmen dieser Simulationen modelliere ich Ausreißer ebenfalls durch lineare Regressionsverteilungen, allerdings mit wesentlich weniger Punkten und sehr großer Störvarianz. Man kann die erzeugten Datensätze also als Datensätze mit $s-1$ Clustern plus Ausreißern interpretieren, wobei die Konzentration auf der Erkennung der $s-1$ „guten" Cluster liegt. Man könnte sich auch für die Schätzung aller s Cluster interessieren, wobei der letzte wesentlich kleiner ist als die anderen. Das soll aber im Rahmen dieser Simulation nicht geschehen.
  - Ein großer Cluster, stark verstreute Ausreißer (out): $s=2$, $p=1,4$, G_1 und Regressionsparameter wie in Situation rand, $G_2=\mathcal{N}_{(0,100\mathrm{I}_p)}$, $n(2)=\frac{n(1)}{5}$, $\sigma_1^2=0.01$, $\sigma_2^2=100$, $n(1)=20,50,100,300$.
  - Zwei große Cluster, stark verstreute Ausreißer (randout): $s=3$, G_1,G_2 und Regressionsparameter wie in Situation rand, $G_3=\mathcal{N}_{(0,100\mathrm{I}_p)}$, $n(3)=\frac{n(1)}{10}$, $\sigma_3^2=100$, $n(1)=n(2)=20,50,100,300$, wobei zwei Situationen simuliert werden: $p=1$, $\sigma_1^2=\sigma_2^2=0.01$ und $p=4$, $\sigma_1^2=\sigma_2^2=0.001$.
  - Zwei große Cluster mit unterschiedlichen Regressorenverteilungen, Ausreißer mit größerer Varianz (randoutx): $s=3$, $p=2$, $\sigma_1^2=\sigma_2^2=0.01$, G_1,G_2,G_3 und Regressionsparameter wie in Situation randx; $n(3)=\frac{n(1)}{10}$, $\sigma_3^2=9$, $n(1)=n(2)=20,50,100,300$.

Insgesamt sind das 110 Konstellationen, die jeweils für alle drei Verfahren simuliert werden. Die Anzahl der Durchläufe pro Simulation war 1000.[25]

Im folgenden werden für die Konstellationen Kurzbezeichnungen nach folgender Nomenklatur (von links nach rechts) verwendet:

- Die Dimension p (bei 1randII3 also $p=1$),
- der Mechanismus zur Erzeugung der Regressoren und Parameter (bei 1randII3 also rand),
- $-\log_{10}$ der Störvarianz (römische Zahl) des größten Clusters (bei 1randII3 ist also II entscheidend, die Störvarianz ist $10^{-2}=0.01$) und
- die Anzahl der Cluster (bei 1randII3 also 3).

15.3 Die erhobenen Statistiken

Ausgegeben werden in jeder Simulation folgende Statistiken:

- Die Häufigkeitsverteilung der Anzahlen gefundener (FPCA) bzw. geschätzter Cluster sowie deren Mittelwert,
- Die Anzahl der Simulationsläufe, in denen jeder modellseitig vorhandene Cluster mindestens einmal „korrekt gefunden" wird nach jedem der drei im folgenden erklärten Kriterien. Ein modellseitig vorhandener Cluster j, $j=1,\ldots,s$ ist charakterisiert durch seinen Regressionsparameter β_j, seinen Skalenparameter σ_j^2, die

[25] Einige der ersten Simulationen hatten 5000 Durchläufe. Da überall relative Häufigkeiten tabelliert sind, ist das aber für die Vergleichbarkeit der Werte unproblematisch.

Indexmenge

$$M_j = \left\{ i \in \{1, \ldots, n\} : \gamma(i) = (\beta_j, \sigma_j^2) \right\}$$

der Punkte, die nach diesen Parametern erzeugt wurden ($\mathbf{X}_{M_j}$ bezeichne die Matrix der entsprechenden Regressorenpunkte) sowie die Regressorenverteilung G_j. Die Schätzung von G_j ist aber nicht von Interesse, sie wird nicht zur Definition der Kriterien herangezogen. $\hat{\beta}, \hat{\sigma}^2, \hat{M}$ seien die Schätzungen der entsprechenden Parameter durch die Verfahren. Dabei ist für die ML-Verfahren

$$\hat{M}_k := \{ i \in \{1, \ldots, n\} : \hat{\zeta}(i) = k \}, \; k = 1, \ldots, \hat{s}$$

und für die FPCA $\hat{M}(g) := \{ i \in \{1, \ldots, n\} : g_i = 1 \}$. Daß ein modellseitig vorhandener Cluster „korrekt gefunden" wird, bedeutet, daß das betreffende Verfahren einen Cluster hervorgebracht hat, der eines oder mehrere der folgenden Kriterien erfüllt:

β-Kriterium: Ein geschätzter Cluster erfüllt das β-Kriterium bzgl. (β, σ^2, M), falls

$$\frac{(\hat{\beta} - \beta)' \mathbf{X}_M' \mathbf{X}_M (\hat{\beta} - \beta)}{\sigma^2} \leq \chi^2_{p+1}(0.95),$$

wobei χ^2_{p+1} die Verteilungsfunktion der χ^2-Verteilung mit $p+1$ Freiheitsgraden bezeichnet. Dieses Kriterium erfüllt der KQ-Schätzer, angewendet auf die Daten aus M, mit Wahrscheinlichkeit 0.95 (siehe zum Beispiel Fahrmeir und Hamerle (1984), S. 89). Das heißt, daß 95% der zu erwartende optimal mögliche Anteil von „korrekt gefundenen" Clustern ist, denn die Parameterschätzung innerhalb der komplizierteren Clustermodelle ist natürlich schwieriger, als wenn eine Schätzung nur aufgrund der „richtigen Daten" vorgenommen wird.

$\beta - \sigma^2$-Kriterium: Ein geschätzter Cluster erfüllt das $\beta - \sigma^2$-Kriterium bzgl. (β, σ^2, M), falls er das β-Kriterium erfüllt und

$$\chi^2_{|M|-p-1}(0.025) \leq \frac{(|M| - p - 1)\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{|M|-p-1}(0.975).$$

Dieses Kriterium erfüllt der KQ-Schätzer zusammen mit dem optimal erwartungstreuen Skalenschätzer, angewendet auf die Daten aus M, mit Wahrscheinlichkeit $0.95^2 = 0.9025$ (siehe zum Beispiel Fahrmeir und Hamerle (1984), S. 89).

Zuordnungskriterium: Ein geschätzter Cluster erfüllt das Zuordnungskriterium bzgl. (β, σ^2, M), falls $|M \setminus \hat{M}| \leq \frac{n}{20}$ und $|\hat{M} \setminus M| \leq \frac{n}{20}$ (jeweils abgerundet).

Jeder vom Verfahren gefundene Cluster wird mit jedem modellseitig vorhandenen Cluster verglichen. Zu beachten ist, daß alle Kriterien an n angepaßt sind. Eine Verbesserung der Ergebnisse gemessen in diesen Kriterien ist daher mit Vergrößerung von n nicht unbedingt zu erwarten. Im Falle des $\beta-$ und $\beta - \sigma^2-$Kriteriums ist nur zu erwarten, daß die Werte mit steigendem n nicht fallen, wenn die „Konvergenzgeschwindigkeit" der Parameterschätzer dieselbe ist wie bei der KQ-Schätzung mit optimaler Varianzschätzung. Verbessern können sich die Ergebnisse nur dann, wenn die allgemeine Konstellation bei steigendem n besser erkannt wird.

16 Simulationsergebnisse

Die Diskussion der Ergebnisse ist nach den in Abschnitt 15.2 eingeführten Konstellationen geordnet.

16.1 Homogene Populationen

Im Falle homogener Populationen interessiert bei den ML-Verfahren nur, wie häufig die korrekte Clusterzahl 1 geschätzt wird. In diesem Fall ist die Schätzung der Parameter die übliche optimale Schätzung. Anderenfalls wird die Konstellation der Daten falsch eingeschätzt, wobei ich es unerheblich finde, ob nun 3 oder 6 Cluster geschätzt werden. Die Beurteilung der FPCA ist etwas komplizierter, weil auch dann ein Cluster vorhanden sein kann, der dem Gesamtdatensatz entspricht, wenn mehr als ein FPC gefunden wird. Die folgenden Tabellen enthalten die relativen Häufigkeiten über alle Simulationsläufe dafür, daß die Clusterzahl 1 geschätzt wurde. Zusätzlich wird die relative Häufigkeit dafür angegeben, daß die modellseitigen Parameter nach dem $\beta - \sigma^2$-Kritierium gefunden werden. Die Ergebnisse nach dem β−Kriterium sind in Bezug auf den Vergleich der Verfahren praktisch identisch zu denen des $\beta - \sigma^2$−Kriteriums. Das Zuordnungskriterium ist erfüllt, wenn es einen Cluster gibt, der ungefähr den ganzen Datensatz enthält. Das ist bei der FPCA immer der Fall und bei den ML-Verfahren genau dann, wenn die korrekte Clusterzahl 1 geschätzt wird.

hom	FPCA: 1 Cluster gefunden				FPCA: $\beta - \sigma^2$-Kriterium			
$p =$	1	2	4	9	1	2	4	9
$n = 20$	0.326	0.131	0.090		0.911	0.912	0.907	
$n = 50$	0.563	0.283	0.006	0	0.899	0.907	0.899	0.898
$n = 100$	0.788	0.634	0.186	0	0.892	0.896	0.896	0.888
$n = 300$	0.875	0.843	0.686	0.110	0.894	0.884	0.893	0.891

hom	MML: 1 Cluster gefunden				MML: $\beta - \sigma^2$−Kriterium			
$p =$	1	2	4	9	1	2	4	9
$n = 20$	0.600	0.190	0.042		0.581	0.200	0.051	
$n = 50$	0.930	0.664	0.312	0.029	0.867	0.627	0.285	0.029
$n = 100$	0.977	0.953	0.919	0.610	0.892	0.877	0.841	0.544
$n = 300$	0.981	0.931	0.970	0.920	0.905	0.902	0.902	0.833

hom	FPML: 1 Cluster gefunden				FPML: $\beta - \sigma^2$-Kriterium			
$p =$	1	2	4	9	1	2	4	9
$n = 20$	0.947	0.904	0.705		0.856	0.820	0.643	
$n = 50$	1	0.999	1	0.969	0.902	0.897	0.899	0.854
$n = 100$	1	1	1	1	0.904	0.906	0.905	0.901
$n = 300$	1	1	1	1	0.904	0.899	0.902	0.896

Die Ergebnisse für die FPCA belegen die Relevanz von Satz 12.1 für Datensätze: Es existiert offenbar für beliebiges n, p fast immer ein FPC, dessen Parameterschätzung

annähernd so gut ist wie der KQ-Schätzer mit optimaler Varianzschätzung. Mit steigendem n wird immer häufiger nur ein Cluster gefunden, was asymptotische Eindeutigkeit vermuten läßt. Die MML-Schätzung überschätzt für kleines n häufig die Clusterzahl, scheint aber auch asymptotisch akzeptabel zu sein. Die FPML-Schätzung funktioniert in dieser Konstellation ausgezeichnet.

Wie sich in den folgenden Abschnitten herausstellen wird, findet die FPCA in den meisten anderen Konstellationen deutlich mehr Cluster. Man kann die Anzahl der gefundenen Fixpunktcluster als Teststatistik für die Homogenitätshypothese verwenden. Aus diesem Grund folgt nun noch eine Tabelle für die durchschnittliche Clusterzahl und die 0.95- bzw. 0.99-Quantile der empirischen Verteilung der gefundenen Clusterzahlen. Die aus der Tabelle ersichtliche Tendenz, daß bei hohem p und kleinem n häufig sehr viele Cluster gefunden werden, wird sich in den weiteren Ergebnissen fortsetzen.

FPCA	Clusteranzahl											
hom	durchschnittlich				0.95-Quantil				0.99-Quantil			
$p=$	1	2	4	9	1	2	4	9	1	2	4	9
$n=20$	2.26	3.37	4.12		5	7	8		6	9	11	
$n=50$	1.59	2.32	6.51	44.70	3	5	11	63	4	6	13	67
$n=100$	1.25	1.48	2.66	13.86	2	3	5	20	3	4	6	23
$n=300$	1.14	1.18	1.39	3.06	2	2	3	6	3	3	4	7

16.2 Konstellationen mit festen Parameterwerten

Für die Konstellationen cross, par und lok ist $s = 2$, und beide Cluster sind aufgrund der symmetrischen Konstellation gleich zu behandeln. Daher wurden in den Tabellen darüber, wie häufig die Cluster gefunden werden, die Werte für beide Cluster zusammengenommen. Tabelliert ist wieder die relative Häufigkeit über alle Simulationsläufe. Bei der Schätzung der Zahl der Cluster ist die relative Häufigkeit für die Schätzung der korrekten Clusterzahl 2 für die ML-Verfahren tabelliert. Da es immer einen FPC gibt, der fast dem kompletten Datensatz entspricht (siehe die Diskussion in Abschnitt 10), wird damit die relative Häufigkeit von drei gefundenen Clustern der FPCA verglichen.

Auch in den folgenden Abschnitten wird die korrekte Clusterzahl+1 als „korrekte Clusterzahl" für die FPCA gewertet.

Außerdem ist die mittlere geschätzte bzw. gefundene Clusterzahl tabelliert.

cross	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1)=20$	13.30	2.13	2	0	0.878	1
$n(1)=50$	13.26	2.02	2	0.001	0.985	1
$n(1)=100$	12.41	2	2	0.002	0.996	1
$n(1)=300$	10.10	2	2	0.058	0.997	1

cross	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	0.896	0.900	0.928	0.704	0.800	0.862	0.511	0.676	0.934
$n(1) = 50$	0.870	0.939	0.936	0.510	0.874	0.864	0.314	0.843	0.987
$n(1) = 100$	0.848	0.936	0.936	0.318	0.871	0.864	0.170	0.890	0.999
$n(1) = 300$	0.808	0.938	0.935	0.142	0.875	0.837	0.027	0.872	1

par	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	5.51	2.86	1.99	0.093	0.471	0.992
$n(1) = 50$	4.10	2.36	1.98	0.328	0.701	0.980
$n(1) = 100$	3.69	2.26	1.99	0.515	0.773	0.994
$n(1) = 300$	3.42	2.19	2	0.655	0.827	0.998

par	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	0.884	0.790	0.929	0.836	0.583	0.871	0.938	0.574	0.990
$n(1) = 50$	0.923	0.906	0.921	0.866	0.797	0.873	0.982	0.839	0.981
$n(1) = 100$	0.936	0.922	0.939	0.870	0.837	0.887	0.994	0.905	0.994
$n(1) = 300$	0.938	0.929	0.947	0.866	0.866	0.890	0.997	0.946	0.999

lok	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	2.97	2.24	1.99	0.264	0.749	0.993
$n(1) = 50$	2.23	2.36	2	0.177	0.673	1
$n(1) = 100$	1.71	2.50	2	0.112	0.568	1
$n(1) = 300$	1.11	2.87	2	0.016	0.356	1

lok	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	0.629	0.872	0.927	0.578	0.757	0.860	0.631	0.819	0.982
$n(1) = 50$	0.416	0.878	0.935	0.393	0.762	0.867	0.436	0.840	0.999
$n(1) = 100$	0.252	0.851	0.942	0.234	0.728	0.875	0.266	0.802	1
$n(1) = 300$	0.041	0.763	0.937	0.037	0.639	0.863	0.044	0.715	1

Zwischen den Konstellationen cross und par ist trotz $p = 2$ und $s = 2$ in beiden Situationen ein deutlicher Unterschied zu sehen.

In cross überschneiden sich die Cluster, so daß insbesondere die FPCA viel größere Schwierigkeiten hat, sie zu trennen. Bei der Zuordnung und der Schätzung von σ^2

ist das Verfahren unbrauchbar, einzig die Schätzung von β ist zufriedenstellend. Dieses Ergebnis steht insofern in Einklang mit Beispiel 13.14, als daß dort die Schranken für $\sigma^2(g, P)$ auch sehr großzügig sind. Dort wird nur die Existenz von FPC gezeigt, deren β nahe dem Modellparameter liegt. Solche FPC werden auch hier gefunden. In der Praxis müssen diese Cluster leider in einer Ausgabe von durchschnittlich mehr als zehn Fixpunktclustern gefunden werden. Dabei ist allerdings zu beachten, daß die FPC erfahrungsgemäß meistens Obermengen der Modellcluster sind, so daß man durch sorgfältige Analyse der Ausgabe durchaus die richtige Konstellation erkennen kann (siehe auch die Analyse des artifiziellen Datensatzes in Abschnitt 10). MML und FPML liefern für cross gute Resultate.

In Situation par dagegen findet die FPCA die Cluster überraschenderweise sogar bzgl. der Parameterschätzung besser als MML. MML hat im Gegensatz zu cross hier die Tendenz, die Clusterzahl zu überschätzen und hinterläßt bei kleinem n einen schlechten Eindruck.

In Situation lok schließlich ist offenbar der Abstand zwischen beiden Clustern so klein, daß das Ergebnis aus Beispiel 13.6 nicht mehr übertragbar ist. Mit wachsendem n tauchen so viele Punkte in der Überschneidung der Cluster auf, daß beide immer häufiger in einen FPC zusammenfallen, so daß die FPCA nur noch einen Cluster findet. MML überschätzt dagegen mit wachsendem n die Clusterzahl erstaunlicherweise immer stärker.

FPML arbeitet in allen diesen Konstellationen ausgezeichnet und ist durchweg besser als beide anderen Verfahren.

Die Konstellation id ist etwas anders geartet als die anderen Situationen. Die korrekte Lösung für die Clusterzahl wäre eigentlich 6, aufgrund des Identifizierbarkeitsproblems kann der Gesamtdatensatz aber mit nur 3 Clustern nahezu perfekt angepaßt werden. Bei der Clusterzahl wird also die relative Häufigkeit für 3 (bzw. 4 bei FPCA) und 6 (bzw. 7 bei FPCA) neben der durchschnittlichen Anzahl gefundener Cluster ausgegeben. Die Cluster sind nicht symmetrisch. Das Simulationsprogramm gibt die Findungshäufigkeiten der Modellcluster aus, nicht die der eigentlich interessanten echten Cluster[26]. Ich beschränke mich hier auf das β–Kriterium. Die Findungshäufigkeiten der Modellcluster mit den Indizes 1,2,9 („einfache Cluster") sind zusammengefaßt, ebenso diejenigen der Modellcluster mit den Indizes 3,4,6, deren Regressionsparameter mit doppelt so vielen Punkten vertreten sind („doppelte Cluster"): Modellcluster 5,7,8 werden mit denselben Regressionsparametern erzeugt.

Nach dem Zuordnungskriterium finden die Verfahren praktisch keinen Modellcluster. Die Ergebnisse des $\beta - \sigma^2$-Kriteriums liegen für alle Verfahren ähnlich; zumindest die einfachen Cluster werden kaum gefunden. Das ist keine Überraschung, da die Cluster so definiert sind, daß jeder der neun Modellcluster von den Regressionsparametern je zweier unterschiedlicher echter Cluster gut angepaßt wird. Daher ist hier nur die Schätzung der Regressionsparameter der echten Cluster von Interesse.

[26] Zur Unterscheidung in „Modell-" und „echte Cluster" siehe die Beschreibung der Konstellation id auf Seite 144.

id	Clusteranzahl								
	durchschnittlich			3 bzw. 4			6 bzw. 7		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
$n(1)=20$	12.07	4.59	3.03	0	0.024	0.966	0.005	0.130	0
$n(1)=50$	11.56	4.75	3.06	0	0.064	0.933	0.007	0.241	0

id	β-Kriterium					
	einfache Cluster			doppelte Cluster		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1)=20$	0.967	0.623	0.745	0.952	0.743	0.514
$n(1)=50$	0.928	0.489	0.541	0.946	0.592	0.662

Die Unterteilung in einfache und doppelte Cluster zeigt nicht viel; eine genaue Analyse der Findungshäufigkeiten aller 9 Modellcluster hätte einiges über die Bedeutung der Lage der Punkte in dieser Konstellation aussagen können, aber darum geht es hier nicht. Alles wesentliche über den Vergleich der Verfahren ist schon hier zu sehen. Die FPCA ist den anderen Verfahren durch den Verzicht auf Zwangspartitionierung überlegen und kann daher alle Cluster trotz der Überschneidungen finden. Bei der Clusteranzahl schätzt FPML üblicherweise drei Cluster, MML überraschenderweise häufiger vier oder fünf als drei oder sechs. Die FPCA findet etwas mehr als die sechs echten Cluster.

16.3 Gleichartige Cluster mit zufälligen Regressionsparametern

In diesem Abschnitt sind wieder alle Cluster symmetrisch, daher werden die Findungshäufigkeiten aller Cluster zusammengerechnet.

16.3.1 Alle Regressorenverteilungen gleich

Zuerst werden die Konstellationen rand behandelt, in denen alle Regressorenverteilungen gleich sind. Die Tabellierung beginnt mit den Situationen mit $s=2$:

rand	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
1randII2	$p=1,\ \sigma_1^2=0.01$					
$n(1)=20$	7.42	2.02	1.97	0.075	0.941	0.972
$n(1)=50$	6.51	2	1.97	0.133	0.977	0.974
$n(1)=100$	5.67	2	1.98	0.171	0.985	0.975
$n(1)=300$	4.55	2.01	1.98	0.205	0.982	0.981
4randIII2	$p=4,\ \sigma_1^2=0.001$					
$n(1)=20$	24.82	2.08	2	0	0.924	1
$n(1)=50$	17.02	2	2	0.003	1	1
$n(1)=100$	12.09	2	2	0.009	1	1
$n(1)=300$	7.49	2	2	0.058	1	1

rand	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1randII2	$p = 1,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.810	0.910	0.902	0.667	0.833	0.835	0.577	0.731	0.868
$n(1) = 50$	0.730	0.926	0.907	0.530	0.864	0.836	0.495	0.759	0.894
$n(1) = 100$	0.663	0.926	0.912	0.416	0.865	0.833	0.452	0.774	0.912
$n(1) = 300$	0.594	0.924	0.910	0.286	0.869	0.790	0.439	0.732	0.927
4randIII2	$p = 4,\ \sigma_1^2 = 0.001$								
$n(1) = 20$	0.912	0.941	0.937	0.815	0.940	0.881	0.960	0.941	0.998
$n(1) = 50$	0.934	0.947	0.948	0.829	0.938	0.895	0.988	0.987	1
$n(1) = 100$	0.931	0.954	0.954	0.808	0.945	0.907	0.995	0.997	1
$n(1) = 300$	0.932	0.947	0.939	0.752	0.932	0.885	0.995	1	1

Es folgen die Resultate mit drei Clustern:

rand	Clusteranzahl					
	durchschnittlich			korrekt (3 bzw. 4)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
1randI3	$p = 1,\ \sigma_1^2 = 0.1$					
$n(1) = 20$	2.76	2.55	4.12	0.172	0.441	0.179
$n(1) = 50$	2.20	2.67	4.81	0.201	0.612	0.025
$n(1) = 100$	2.04	2.77	5.05	0.081	0.696	0.012
$n(1) = 300$	1.85	2.90	5.33	0.036	0.808	0
4randIII3	$p = 4,\ \sigma_1^2 = 0.001$					
$n(1) = 20$	12.19	3.32	2.96	0.008	0.747	0.960
$n(1) = 50$	8.16	3.02	3	0.074	0.980	1
$n(1) = 100$	6.80	3.02	3	0.145	0.981	1
$n(1) = 300$	5.61	3.01	3	0.236	0.989	1
9randIII3	$p = 9,\ \sigma_1^2 = 0.001$					
$n(1) = 50$	9.42	3.16	2.97	0.036	0.850	0.954
$n(1) = 150$	3.69	3.01	3	0.164	0.993	1

rand	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1randl3	$p = 1,\ \sigma_1^2 = 0.1$								
$n(1) = 20$	0.155	0.576	0.530	0.078	0.457	0.279	0.032	0.131	0.148
$n(1) = 50$	0.054	0.657	0.340	0.023	0.549	0.172	0.018	0.121	0.080
$n(1) = 100$	0.028	0.684	0.250	0.013	0.580	0.134	0.015	0.129	0.072
$n(1) = 300$	0.011	0.728	0.159	0.006	0.615	0.087	0.010	0.115	0.043
4randlll3	$p = 4,\ \sigma_1^2 = 0.001$								
$n(1) = 20$	0.216	0.924	0.884	0.145	0.922	0.830	0.264	0.890	0.953
$n(1) = 50$	0.465	0.940	0.940	0.319	0.934	0.885	0.505	0.979	1
$n(1) = 100$	0.619	0.945	0.940	0.407	0.934	0.881	0.690	0.995	1
$n(1) = 300$	0.741	0.936	0.939	0.360	0.918	0.886	0.822	0.996	1
9randlll3	$p = 9,\ \sigma_1^2 = 0.001$								
$n(1) = 50$	0.017	0.942	0.897	0.011	0.939	0.848	0.022	0.959	0.966
$n(1) = 150$	0.091	0.938	0.935	0.058	0.933	0.875	0.101	0.999	1

Nun noch die Ergebnisse mit vier Clustern:

rand	Clusteranzahl					
	durchschnittlich			korrekt (4 bzw. 5)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
1randlll4	$p = 1,\ \sigma_1^2 = 0.001$					
$n(1) = 20$	6.69	4.18	3.83	0.143	0.750	0.838
$n(1) = 50$	5.87	4.16	3.83	0.186	0.793	0.832
$n(1) = 100$	5.40	4.15	3.85	0.193	0.815	0.849
$n(1) = 300$	4.82	4.13	3.88	0.204	0.836	0.883

rand	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1randlll4	$p = 1,\ \sigma_1^2 = 0.001$								
$n(1) = 20$	0.439	0.903	0.854	0.282	0.892	0.794	0.369	0.865	0.908
$n(1) = 50$	0.460	0.916	0.857	0.263	0.898	0.794	0.417	0.894	0.912
$n(1) = 100$	0.468	0.916	0.866	0.223	0.896	0.799	0.441	0.904	0.923
$n(1) = 300$	0.451	0.918	0.881	0.132	0.895	0.785	0.460	0.915	0.940

In dieser Situation ist das MML-Verfahren im Vorteil, denn gegeben den Regressor x sind, wie im Mischmodell 1 gefordert, die Mischungsproportionen immer gleich, d.h. die Zuordnungsunabhängigkeit ist erfüllt. Diese Situation wird für das MML-Verfahren vorausgesetzt. Entsprechend erzielt MML von allen Verfahren die besten Ergebnisse bei den Parameterschätzungskriterien β und $\beta - \sigma^2$. Interessant ist auch, daß die Verluste gegenüber der KQ-Schätzung mit optimaler Skalenschätzung sehr klein sind. In den Situationen mit sehr kleiner Störvarianz (0.001) erreicht MML sogar häufig bessere Werte als die theoretisch zu erwartende 0.9025. Das liegt daran, daß innerhalb der MML-

Berechnung der Mindestwert für σ^2 auf 0.001 festgelegt ist, so daß in diesen Situationen das $\beta - \sigma^2$-Kriterium nicht durch Unterschätzung der Störvarianz verletzt werden kann. Es handelt sich also um ein Artefakt der Simulation. Auch im folgenden ist bei allen Konstellationen mit $\sigma_i^2 = 0.001$ für einen Cluster zu beachten, daß die Ergebnisse von MML bzgl. des $\beta-$ und insbesondere des $\beta-\sigma^2-$Kriteriums besser sind, als sie bei Wahl einer anderen MML-Varianzuntergrenze wären. Die Schätzung der Clusterzahl durch das BIC ist in allen Fällen gut.

Die explizite Schätzung der Partition des Datensatzes im Fixed Partition Model (Modell 2) äußert sich darin, daß FPML auch in diesem Abschnitt trotz optimaler Bedingungen für MML nach dem Zuordnungskriterium meistens besser abschneidet. Die Parameterschätzungen sind zufriedenstellend, die Schätzung der Clusterzahl ausgezeichnet. Eine Ausnahme davon bildet die Situation 1randl3 (siehe Abbildung 10). In dieser Konstellation bricht das Verfahren zusammen. Offenbar benötigt FPML ähnlich wie die FPCA gut voneinander getrennte Cluster. Etwas überraschend ist jedoch, daß FPML hier dazu neigt, die Clusterzahl zu überschätzen, im Gegensatz zu den meisten anderen Konstellationen und zur FPCA. Der größte Vorteil von MML scheint die Überlegenheit in der Konstellation 1randl3 zu sein: In Datensätzen, deren Struktur stark vom Störterm überlagert wird, kann nur ein Verfahren bestehen, für das die Struktur weitgehend vorausgesetzt wird. Die Zuordnungsschätzung, allgemein die Schwäche von MML, ist in dieser Situation ohnehin nicht gut möglich. Das kann man an den Ergebnissen aller Verfahren beim Zuordnungskriterium sehen.

Das Abschneiden der FPCA in den rand-Konstellationen ist allgemein schlecht: In einigen Konstellationen (1randl3, 9randlll2) werden kaum korrekte Cluster gefunden. Bei 1randll2 verschlechtern sich die Ergebnisse mit steigendem n. In allen Konstellationen außer 4randlll2 sind die Findungshäufigkeiten deutlich niedriger als die der anderen Verfahren, insbesondere nach dem $\beta - \sigma^2-$Kriterium. Die Schätzung von σ^2 ist offenbar immer stark verzerrt.

Das Verhalten der FPCA kann zwei Gründe haben: Bei 1randl3 und eventuell 1randll2 sind die Cluster nicht gut genug getrennt; die FPCA faßt Punkte aus unterschiedlichen Modellclustern zusammen. In den Fällen mit $p = 9$ oder $s = 4$ reicht vermutlich die Iterationszahl nicht aus, um die relevanten FPCV zu finden.

16.3.2 Unterschiedliche Regressorenverteilungen

In der Situation randlx liegen die Regressoren für die verschiedenen Cluster mit einer Wahrscheinlichkeit von nahezu 1 in jeder Dimension so weit auseinander, daß das bei einer optischen Vorabanalyse des Datensatzes sofort auffiele. Es scheint mir nicht besonders realistisch zu sein, einen solchen Datensatz zu analysieren, ohne daß vorher die offensichtlich verschiedenartigen Daten auseinandersortiert werden. Diese Konstellation wird nur simuliert, um zu sehen, wie sich unterschiedliche Regressoren im Extremfall auswirken. Daher beschränke ich mich auf jeweils einen Wert von n. Die Tabellierung beginnt wieder mit $s = 2$:

	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
change	$p=1,\ \sigma_1^2=0.01$					
$n(1)=20$	4.21	2.13	1.78	0.248	0.320	0.702
$n(1)=50$	3.08	2.03	1.81	0.328	0.437	0.661
$n(1)=100$	2.68	2.18	1.82	0.325	0.524	0.676
$n(1)=300$	2.44	2.72	1.84	0.326	0.375	0.698
randx						
2randxII2	$p=2,\ \sigma_1^2=0.01$					
$n(1)=20$	13.02	2.48	1.99	0.013	0.653	0.858
$n(1)=50$	10.54	2.18	2	0.055	0.818	0.894
$n(1)=100$	8.87	2.16	2	0.108	0.857	0.921
$n(1)=300$	6.88	2.15	2	0.166	0.879	0.968
4randxIII2	$p=4,\ \sigma_1^2=0.001$					
$n(1)=20$	29.61	3.17	1.97	0.002	0.502	0.855
$n(1)=50$	22.55	2.19	2.03	0.033	0.846	0.896
$n(1)=100$	16.81	2.08	2.01	0.042	0.920	0.962
$n(1)=300$	9.49	2.06	1.99	0.098	0.961	0.989
rand!x						
2rand!xII2	$p=2,\ \sigma_1^2=0.01$					
$n(1)=100$	9.47	2.52	2.12	0.202	0.504	0.606
9rand!xIII2	$p=9,\ \sigma_1^2=0.001$					
$n(1)=150$	29.02	2.32	2.08	0.046	0.783	0.813

	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
change	$p = 1,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.750	0.436	0.655	0.657	0.316	0.580	0.617	0.143	0.523
$n(1) = 50$	0.617	0.350	0.628	0.549	0.272	0.562	0.577	0.136	0.506
$n(1) = 100$	0.556	0.313	0.631	0.502	0.251	0.555	0.555	0.132	0.539
$n(1) = 300$	0.538	0.227	0.596	0.501	0.164	0.510	0.562	0.112	0.568
randx									
2randx||2	$p = 2,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.921	0.711	0.803	0.845	0.604	0.736	0.868	0.537	0.804
$n(1) = 50$	0.867	0.820	0.843	0.765	0.744	0.779	0.811	0.714	0.866
$n(1) = 100$	0.825	0.856	0.871	0.688	0.790	0.806	0.784	0.776	0.906
$n(1) = 300$	0.737	0.874	0.900	0.572	0.805	0.822	0.754	0.815	0.946
4randx|||2	$p = 4,\ \sigma_1^2 = 0.001$								
$n(1) = 20$	0.892	0.620	0.775	0.841	0.618	0.728	0.978	0.551	0.836
$n(1) = 50$	0.932	0.849	0.869	0.882	0.844	0.821	0.999	0.865	0.910
$n(1) = 100$	0.932	0.902	0.914	0.874	0.894	0.856	0.996	0.927	0.968
$n(1) = 300$	0.899	0.933	0.936	0.797	0.915	0.885	0.999	0.987	0.990
rand!x									
2rand!x||2	$p = 2,\ \sigma_1^2 = 0.01$								
$n(1) = 100$	0.912	0.376	0.632	0.833	0.326	0.545	0.947	0.303	0.663
9rand!x|||2	$p = 9,\ \sigma_1^2 = 0.001$								
$n(1) = 150$	0.522	0.332	0.559	0.490	0.329	0.528	0.732	0.418	0.664

Es folgen die Resultate mit drei Clustern:

randx	Clusteranzahl					
	durchschnittlich			korrekt (3 bzw. 4)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
1randx||3	$p = 1,\ \sigma_1^2 = 0.01$					
$n(1) = 20$	8.80	3.15	2.54	0.061	0.402	0.567
$n(1) = 50$	7.33	3.31	2.59	0.108	0.463	0.605
$n(1) = 100$	6.50	3.44	2.61	0.119	0.494	0.623
$n(1) = 300$	5.52	3.65	2.64	0.157	0.467	0.650
2randx|||3	$p = 2,\ \sigma_1^2 = 0.001$					
$n(1) = 20$	14.90	3.56	2.93	0.008	0.388	0.862
$n(1) = 50$	12.23	3.37	3.01	0.017	0.598	0.876
$n(1) = 100$	10.40	3.32	3.01	0.034	0.691	0.907
$n(1) = 300$	8.17	3.26	3.02	0.070	0.760	0.952

randx	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1randxII3	$p = 1$, $\sigma_1^2 = 0.01$								
$n(1) = 20$	0.629	0.607	0.654	0.496	0.503	0.587	0.508	0.404	0.629
$n(1) = 50$	0.544	0.662	0.662	0.409	0.569	0.601	0.469	0.475	0.668
$n(1) = 100$	0.493	0.683	0.666	0.355	0.585	0.600	0.476	0.522	0.701
$n(1) = 300$	0.423	0.655	0.675	0.278	0.546	0.585	0.463	0.526	0.733
2randxIII3	$p = 2$, $\sigma_1^2 = 0.001$								
$n(1) = 20$	0.541	0.591	0.804	0.457	0.586	0.752	0.706	0.572	0.872
$n(1) = 50$	0.657	0.776	0.872	0.570	0.766	0.816	0.788	0.770	0.927
$n(1) = 100$	0.693	0.823	0.888	0.576	0.811	0.834	0.834	0.840	0.952
$n(1) = 300$	0.717	0.862	0.920	0.544	0.841	0.860	0.857	0.888	0.982

Nun noch die Ergebnisse mit vier Clustern:

randx	Clusteranzahl					
	durchschnittlich			korrekt (4 bzw. 5)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
4randxIII4	$p = 4$, $\sigma_1^2 = 0.001$					
$n(1) = 20$	12.61	3.91	2.01	0.029	0.182	0.090
$n(1) = 50$	7.27	3.99	2.71	0.110	0.303	0.274
$n(1) = 150$	5.81	4.35	3.74	0.124	0.555	0.647

randx	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
4randxIII4	$p = 1$, $\sigma_1^2 = 0.001$								
$n(1) = 20$	0.026	0.036	0.082	0.019	0.035	0.074	0.104	0.070	0.126
$n(1) = 50$	0.053	0.350	0.303	0.044	0.346	0.284	0.169	0.430	0.375
$n(1) = 150$	0.117	0.711	0.715	0.092	0.700	0.667	0.254	0.819	0.806

Die Ergebnisse unterscheiden sich deutlich von den Resultaten mit gleichen Regressorenverteilungen. Insbesondere bei kleinem n findet MML wenige korrekte Cluster. Bei change und 1randxII3 wird aber anscheinend auch asymptotisch die Clusterzahl überschätzt. In den rand!x-Konstellationen wird ebenfalls deutlich, daß das Verfahren bei Verletzung der Zuordnungsunabhängigkeit ernsthafte Probleme bekommt, wenn es auch in den realistischeren Konstellationen nicht völlig versagt. Die schlechten Ergebnisse bei change sind etwas überraschend, weil die Regressoren für die einzelnen Cluster dabei im Mittel nicht weiter auseinanderliegen als bei randx. Die Wechselpunkt-Konstellation scheint auch für die anderen Verfahren einen besonderen Schwierigkeitsgrad zu haben, aber MML hat die größten Probleme.

Die FPCA schneidet dagegen deutlich besser ab als im Fall gleicher Regressorenverteilungen. Das Verfahren kann fast durchweg mit den parametrischen Methoden mithalten. Insbesondere die guten Ergebnisse bei rand!x waren nicht unbedingt zu erwarten, be-

denkt man die Voraussetzung an die Regressorenverteilungen in Satz 13.11. Einzig bei $p = 4, s = 4$ hinterläßt das Verfahren einen sehr schwachen Eindruck. In Bemerkung 15.1 wird damit übereinstimmend eine geringe Wahrscheinlichkeit berechnet, vorhandene - besonders kleine - Fixpunktcluster bei hohem p schnell zu finden. Das scheint mir die Achillesferse des Verfahrens zu sein, so wie es bisher implementiert ist. Andererseits schneidet die FPCA in den anderen Konstellationen bei kleinem n mehrfach am besten ab. Das Verfahren scheint gut zur intensiven Analyse kleiner Datensätze geeignet zu sein. Neben den modellseitigen Clustern bekommt man im Schnitt viele weitere. Das kann je nach Ziel der Analyse ein Vor- oder Nachteil sein. Die Clusterzuordnung ist durchweg besser als die Parameterschätzung.

FPML bringt auch in den Konstellationen mit unterschiedlicher Regressorenverteilung gute Ergebnisse. Das Verfahren ist nicht immer das beste, fällt aber gegenüber dem besten nie deutlich ab. Die Clusterzahlschätzung ist fast immer besser als die von MML. und bewegt sich mit steigendem n immer in die richtige Richtung. Im Gegensatz zu MML neigt FPML eher zur Unterschätzung der Clusterzahl.

16.4 Verschiedenartige Cluster

In den Konstellationen mit verschiedenartigen Clustern werden die Findungshäufigkeiten nicht über die Cluster gemittelt, sondern die Ergebnisse sind hier nach den einzelnen Clustern aufgeschlüsselt. Die Clusternummer ist der Index aus Abschnitt 15.2, d.h. der größte Cluster hat immer Index 1.

4halbIII2	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 50$	12.84	2.01	2	0.001	0.990	1
$n(1) = 100$	10.60	2	2	0.020	1	1
$n(1) = 300$	8.96	2	2	0.095	1	1

4halbIII2	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
Cluster 1									
$n(1) = 50$	0.942	0.961	0.940	0.873	0.953	0.879	0.990	0.953	0.996
$n(1) = 100$	0.926	0.948	0.955	0.855	0.939	0.894	0.998	0.991	1
$n(1) = 300$	0.939	0.949	0.940	0.836	0.936	0.888	1	0.997	1
Cluster 2, halbe Größe									
$n(1) = 50$	0.311	0.942	0.945	0.220	0.940	0.896	0.289	0.949	0.996
$n(1) = 100$	0.547	0.948	0.929	0.405	0.941	0.876	0.592	0.991	1
$n(1) = 300$	0.788	0.940	0.942	0.495	0.919	0.892	0.859	0.997	1

2scl2	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	9.80	2.95	2	0.016	0.549	0.994
$n(1) = 50$	7.02	2.07	2	0.105	0.943	1
$n(1) = 100$	5.33	2.02	2	0.183	0.982	1
$n(1) = 300$	3.81	2.02	2	0.275	0.983	1

2scl2	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
Cluster 1, $\sigma_1^2 = 0.1$									
$n(1) = 20$	0.479	0.770	0.926	0.345	0.508	0.876	0.137	0.491	0.929
$n(1) = 50$	0.219	0.925	0.943	0.124	0.854	0.890	0.062	0.839	0.976
$n(1) = 100$	0.121	0.941	0.933	0.052	0.886	0.880	0.047	0.846	0.982
$n(1) = 300$	0.051	0.946	0.935	0.026	0.895	0.878	0.039	0.816	0.988
Cluster 2, $\sigma_1^2 = 0.001$									
$n(1) = 20$	0.943	0.901	0.906	0.863	0.795	0.834	0.934	0.796	0.933
$n(1) = 50$	0.935	0.932	0.934	0.819	0.859	0.865	0.937	0.878	0.976
$n(1) = 100$	0.931	0.940	0.936	0.773	0.876	0.869	0.955	0.853	0.982
$n(1) = 300$	0.918	0.933	0.936	0.663	0.863	0.867	0.957	0.816	0.988

9schalbxll2	Clusteranzahl					
	durchschnittlich			korrekt (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 50$	61.37	3.30	1.88	0	0.452	0.836
$n(1) = 100$	45.39	2.29	1.98	0	0.857	0.927
$n(1) = 300$	31.68	2.03	2	0.055	0.979	0.993

9schalbxll2	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
Cluster 1, $\sigma_1^2 = 0.01$									
$n(1) = 50$	0.968	0.430	0.654	0.931	0.342	0.602	1	0.380	0.720
$n(1) = 100$	0.952	0.846	0.838	0.904	0.734	0.798	0.999	0.858	0.900
$n(1) = 300$	0.933	0.948	0.940	0.846	0.890	0.883	1	0.990	0.992
Cluster 2, halbe Größe, $\sigma_1^2 = 0.001$									
$n(1) = 50$	0.002	0.304	0.592	0.001	0.304	0.568	0.006	0.404	0.728
$n(1) = 100$	0.025	0.803	0.826	0.021	0.800	0.778	0.036	0.885	0.913
$n(1) = 300$	0.232	0.912	0.900	0.206	0.909	0.858	0.327	0.989	0.995

1scvxII3	Clusteranzahl					
	durchschnittlich			korrekt (3 bzw. 4)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML
$n(1) = 20$	10.02	3.31	2.74	0.035	0.415	0.737
$n(1) = 50$	8.56	3.33	2.80	0.076	0.518	0.785
$n(1) = 100$	7.84	3.39	2.83	0.081	0.560	0.805
$n(1) = 300$	6.43	3.50	2.83	0.144	0.573	0.820

1scvxII3	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
Cluster 1, $\sigma_1^2 = 0.01$									
$n(1) = 20$	0.809	0.671	0.783	0.683	0.546	0.726	0.647	0.379	0.701
$n(1) = 50$	0.703	0.746	0.804	0.554	0.641	0.754	0.625	0.567	0.785
$n(1) = 100$	0.661	0.728	0.787	0.494	0.648	0.731	0.628	0.592	0.794
$n(1) = 300$	0.554	0.707	0.811	0.407	0.631	0.744	0.605	0.602	0.839
Cluster 2, $\sigma_1^2 = 0.001$									
$n(1) = 20$	0.730	0.645	0.784	0.655	0.626	0.737	0.699	0.531	0.802
$n(1) = 50$	0.750	0.744	0.799	0.623	0.728	0.746	0.783	0.700	0.855
$n(1) = 100$	0.754	0.803	0.831	0.618	0.773	0.775	0.793	0.739	0.887
$n(1) = 300$	0.734	0.803	0.832	0.554	0.776	0.756	0.823	0.744	0.894
Cluster 3, $\sigma_1^2 = 0.003$									
$n(1) = 20$	0.392	0.604	0.720	0.299	0.590	0.652	0.309	0.476	0.730
$n(1) = 50$	0.389	0.699	0.783	0.289	0.619	0.718	0.374	0.644	0.822
$n(1) = 100$	0.357	0.748	0.767	0.271	0.670	0.716	0.392	0.701	0.827
$n(1) = 300$	0.377	0.756	0.794	0.255	0.666	0.718	0.426	0.705	0.850

Bei MML und FPML unterscheiden sich die Ergebnisse für die verschiedenen Cluster innerhalb einer Konstellation kaum. Zumindest wenn die Clusterzahl richtig geschätzt wurde, ist es fast gleichbedeutend, einen und alle Cluster korrekt zu finden. Im Gegensatz dazu ist die Fähigkeit der FPCA, Cluster zu finden, stark von den Eigenschaften des Clusters im Verhältnis zur Restkonstellation abhängig. Cluster mit großem $n(i)$ und kleinem σ_i^2 werden wesentlich häufiger gefunden. Insbesondere in den hohen Dimensionen bräuchte die FPCA, wie bereits erwähnt und in Bemerkung 15.1 erläutert, für kleine Cluster mehr Iterationen. Aber auch die Ergebnisse für den dritten Cluster in 1scvxII3 sehen nicht gut aus. In 2scI2 werden vermutlich meistens Teile des Clusters 2 mit kleinerer Störskala auch dem ersten Cluster zugerechnet, so daß dieser fast immer verzerrt ist. Bei größerem n werden sogar manchmal nur noch zwei Cluster gefunden, vermutlich der Gesamtdatensatz und der zweite Modellcluster. Wenn die Information vorhanden ist, daß der Datensatz mit einer Partition in mehrere Regressionsverteilungen angemessen zu beschreiben ist, ist dieses Ergebnis aber ausreichend, um die Gesamtkonstellation zu finden. Dann kann der erste Modellcluster einfach ermittelt werden, weil er aus den Punkten besteht, die im Gesamtdatensatz, aber nicht im kleineren Cluster sind. Immerhin findet die FPCA meistens alle bis auf einen Cluster korrekt; manchmal bei kleinem n häufiger als die ML-Verfahren. Man hätte also in der Ausgabe des Verfahrens

alle nötigen Informationen, um sich eine gute Vorstellung von der Gesamtkonstellation zu machen, vorausgesetzt, man würde relevante von irrelevanten Clustern unterscheiden können, wenn zu viele gefunden werden. Bei 9schalbxII2 fällt auf, daß bei kleinem n der größere Cluster nach dem β–Kriterium besser geschätzt wird als beim einfachen KQ-Schätzer theoretisch zu erwarten wäre (0.95). Das ist hier auf die hohe Zahl gefundener Cluster zurückzuführen. Der größere Cluster wird also in mehreren „Versionen" gefunden und es ist nicht überraschend, daß fast immer wenigstens eine dieser Versionen eine gute β–Schätzung enthält.

Der Eindruck bei den ML-Verfahren ähnelt dem aus anderen Konstellationen. MML zeigt deutliche Schwächen bei kleinen Stichproben. FPML ist beim Zuordnungskriterium durchweg überlegen. Lediglich in den Situationen mit gleicher Regressorenverteilung aller Cluster (2scI2 und 4halbIII2) ist die Parameterschätzung manchmal bei MML besser. Dabei ist allerdings der künstliche Vorteil von MML bei 4halbIII2 durch die Störskalenuntergrenze 0.001 zu berücksichtigen. Nur in der Konstellation mit drei Clustern unterschiedlicher Größe und Störskala 1scvxII3 tauchen Probleme bei der Schätzung der Anzahl der Cluster auf. Insbesondere MML schneidet dabei schlecht ab.

16.5 Ausreißerkonstellationen

In den Ausreißerkonstellationen interpretiere ich die Punkte aus dem jeweils kleinsten Cluster als irrelevante Ausreißer. Entscheidend ist also nur, ob die großen Cluster gut gefunden werden. Falls es mehrere sind, sind sie symmetrisch. In den Tabellen für „Cluster korrekt gefunden" werden also die Findungshäufigkeiten über die Nichtausreißercluster gemittelt. Bei den geschätzten Clusterzahlen führe ich sowohl die Häufigkeiten für die Zahl der Nichtausreißercluster als auch für die um 1 höhere Modellclusterzahl an.

out!	Clusteranzahl								
	durchschnittlich			ohne Ausreißer (1 bzw. 2)			mit Ausreißern (2 bzw. 3)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1out!II2	$p=1,\ \sigma_1^2=0.01$								
$n(1)=20$	6.13	3.68	2.04	0.018	0	0.054	0.073	0.021	0.868
$n(1)=50$	4.60	4.66	2.08	0.106	0	0.002	0.174	0.129	0.916
$n(1)=100$	2.84	2.37	2.03	0.263	0	0	0.221	0.799	0.968
$n(1)=300$	1.66	2.04	2	0.237	0	0	0.085	0.971	0.999
4out!II2	$p=4,\ \sigma_1^2=0.01$								
$n(1)=20$	25.82	2.96	1.96	0	0	0.119	0	0.318	0.806
$n(1)=50$	74.46	4.43	2.05	0	0	0.026	0	0.006	0.896
$n(1)=100$	66.58	5.09	2.29	0	0	0.021	0	0.060	0.694
$n(1)=300$	7.70	2.24	2	0.089	0	0	0.113	0.824	0.997

out!	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1out!II2	$p = 1,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.956	0.812	0.800	0.910	0.596	0.716	0.998	0.337	0.811
$n(1) = 50$	0.946	0.832	0.873	0.893	0.726	0.815	1	0.528	0.985
$n(1) = 100$	0.940	0.878	0.885	0.883	0.820	0.827	1	0.708	0.997
$n(1) = 300$	0.930	0.872	0.888	0.874	0.821	0.830	1	0.670	1
4out!II2	$p = 4,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.969	0.520	0.045	0.931	0.272	0.011	1	0.186	0
$n(1) = 50$	0.966	0.704	0.161	0.923	0.456	0.146	1	0.197	0.662
$n(1) = 100$	0.955	0.610	0.517	0.910	0.505	0.486	1	0.562	0.924
$n(1) = 300$	0.931	0.815	0.848	0.883	0.759	0.785	1	0.746	1

randout	Clusteranzahl								
	durchschnittlich			ohne Ausreißer (2 bzw. 3)			mit Ausreißern (3 bzw. 4)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1randoutII3	$p = 1,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	8.77	3.56	2.12	0.045	0.030	0.699	0.059	0.478	0.203
$n(1) = 50$	7.88	4.66	2.78	0.062	0.002	0.235	0.092	0.136	0.703
$n(1) = 100$	6.81	3.82	2.95	0.093	0.009	0.088	0.098	0.525	0.873
$n(1) = 300$	5.32	3.21	2.97	0.120	0.007	0.039	0.124	0.819	0.954
4randoutIII3	$p = 4,\ \sigma_1^2 = 0.001$								
$n(1) = 20$	29.63	3.77	2.14	0	0.002	0.851	0	0.410	0.143
$n(1) = 50$	29.50	4.70	2.66	0	0	0.377	0	0.036	0.588
$n(1) = 100$	25.93	5.36	2.88	0	0	0.250	0.002	0.033	0.626
$n(1) = 300$	16.86	3.63	3.03	0.007	0	0.001	0.020	0.608	0.975

randout	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
1randoutII3	$p = 1,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.805	0.834	0.282	0.660	0.705	0.229	0.565	0.459	0.334
$n(1) = 50$	0.719	0.804	0.626	0.522	0.696	0.555	0.477	0.435	0.681
$n(1) = 100$	0.671	0.840	0.796	0.418	0.767	0.701	0.452	0.428	0.852
$n(1) = 300$	0.574	0.866	0.858	0.269	0.801	0.694	0.419	0.424	0.907
4randoutIII3	$p = 4,\ \sigma_1^2 = 0.001$								
$n(1) = 20$	0.868	0.830	0.230	0.765	0.828	0.200	0.935	0.485	0.504
$n(1) = 50$	0.928	0.860	0.359	0.834	0.858	0.317	0.986	0.641	0.602
$n(1) = 100$	0.931	0.728	0.453	0.809	0.720	0.411	0.991	0.816	0.754
$n(1) = 300$	0.923	0.879	0.911	0.738	0.867	0.848	0.997	0.730	0.999

randoutx	Clusteranzahl								
	durchschnittlich			ohne Ausreißer (2 bzw. 3)			mit Ausreißern (3 bzw. 4)		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
2randoutxII3	$p = 2,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	14.52	3.63	2.07	0.004	0.133	0.780	0.003	0.389	0.140
$n(1) = 50$	11.70	3.85	2.55	0.009	0.044	0.383	0.026	0.394	0.560
$n(1) = 100$	9.64	3.56	2.86	0.024	0.024	0.121	0.065	0.616	0.836
$n(1) = 300$	7.19	3.26	2.95	0.032	0.012	0.056	0.111	0.789	0.923

randoutx	Cluster korrekt gefunden								
Kriterium	β			$\beta - \sigma^2$			Zuordnung		
Verfahren	FPCA	MML	FPML	FPCA	MML	FPML	FPCA	MML	FPML
2randoutxII3	$p = 2,\ \sigma_1^2 = 0.01$								
$n(1) = 20$	0.925	0.581	0.294	0.832	0.441	0.244	0.850	0.286	0.437
$n(1) = 50$	0.854	0.686	0.547	0.750	0.573	0.468	0.814	0.460	0.667
$n(1) = 100$	0.804	0.777	0.768	0.671	0.692	0.667	0.779	0.520	0.874
$n(1) = 300$	0.721	0.806	0.830	0.541	0.725	0.635	0.754	0.510	0.942

Die Ausreißer machen der FPCA am wenigsten zu schaffen. Man vergleiche dazu zum Beispiel die Ergebnisse von 4randoutIII3 mit denen von 4randIII2. Die Findungshäufigkeiten unterscheiden sich kaum. Allerdings wird die Zahl der gefundenen Cluster zumindest bei $p = 4$ sehr groß. Das Verfahren gibt offenbar diverse Punktmengen als Fixpunktcluster aus, in denen ein Teil der Ausreißer mit den Daten aus den großen Clustern zusammenfaßt wird. Besonders bei großem n wird in der Konstellation 1out!II2 andererseits häufig nur noch ein Cluster gefunden. Die Findungshäufigkeiten deuten darauf hin, daß das der korrekte große Cluster ist. Es gibt hier also keinen Cluster mehr, der den ganzen Datensatz inklusive Ausreißern enthält.

Die Ergebnisse der ML-Verfahren sind für große n gut. Der Cluster mit den Ausreißern ist dann groß genug, daß die Verfahren seine eigene Struktur erkennen; die hier simulierten Ausreißer sind ja nicht „modellfremd". Bei nur wenigen Ausreißern bekommen MML und FPML aber Probleme. FPML schätzt dabei häufig die Clusterzahl als die Zahl der Modellcluster ohne Ausreißer. Die Ausreißer werden also in die anderen Cluster „integriert". Das läßt deren Parameterschätzer zusammenbrechen. Die Ergebnisse nach dem Zuordnungskriterium sind nicht ganz so schlecht.

MML hat dagegen die Tendenz, die Clusterzahl zu überschätzen. Offenbar gibt es dann Cluster, die den großen Modellclustern abzüglich einiger Punkte entsprechen, und mehrere kleine Cluster mit Teilen der Ausreißer. Das hat gegenüber FPML den Vorteil, daß in den großen Clustern zumindest die Parameter besser geschätzt werden. Dieses Verhalten ist auch bei den Datensätzen in Abschnitt 10 zu beobachten.

Allgemein unterstützen die Ergebnisse den Verdacht, daß die ML-Verfahren mit Ausreißern Probleme bekommen, die keine sichtbare lineare Struktur haben. Die FPCA macht hier einen deutlich besseren Eindruck, sofern man nicht die große Zahl der gefundenen Cluster als entscheidenden Nachteil wertet.

17 Fazit: Simulationen

Die drei Verfahren, die in den Simulationen verglichen wurden, sind eigentlich für unterschiedliche Situationen gedacht: Der prinzipielle Unterschied zwischen dem Mischmodell 1 und dem Modell 2 mit fester Zuordnung besteht darin, daß ersteres, und damit das MML-Verfahren, die Zuordnungsunabhängigkeit voraussetzt. Der Zweck der FPCA ist eher die Verwendung zur explorativen Datenanalyse, falls man weniger Informationen hat als die anderen Verfahren voraussetzen. Weiterhin unterscheidet sich die Ausgabe der FPCA wesentlich von der der ML-Verfahren.

In den Simulationen wurde künstlich Vergleichbarkeit hergestellt. Trotzdem sind die Ergebnisse relevant, denn die Datenkonstellationen sind so gewählt, daß sich prinzipiell die Verwendung aller Verfahren anbietet. Es ist zum Beispiel nicht damit zu rechnen, daß immer die Zuordnungsunabhängigkeit nachgeprüft werden kann, wenn MML eingesetzt wird. Beim Abschneiden der FPCA ist zu berücksichtigen, daß ich auf Simulationen verzichtet habe, in denen Ausreißer vorkommen, die nicht von einem linearen Regressionsmodell generiert wurden. So bleibt die Vermutung ungeprüft, daß das Verfahren hier besondere Stärken hat.

Die drei Verfahren zeigen erwartungsgemäß ein sehr unterschiedliches Profil. Daher ist die Zusammenfassung der Simulationen nach den Verfahren geordnet.

17.1 Fixpunktclusteranalyse

Vorteile:

- Die FPCA läßt sich am wenigsten von Ausreißern beeinflussen.
- Bei kleinem n, p und s findet die FPCA häufig besser die Cluster als die ML-Verfahren.
- Die FPCA arbeitet gut, wenn die Modellcluster sehr gut voneinander getrennt sind.
- Die FPCA ist in der Lage, unterschiedliche Parameterkonstellationen zu finden, die denselben Datensatz anpassen können. Dadurch können bei Nicht-Identifizierbarkeit der Parameter mehrere Alternativen gefunden werden.

Nachteile:

- Die Anzahl der gefundenen FPC ist zur Schätzung der Zahl der Modellcluster nicht brauchbar. Manchmal werden extrem viele FPC gefunden. Die Ausgabe des Verfahrens wird dadurch sehr unübersichtlich.
- In Situationen mit Clustern, die sich überschneiden (zum Beispiel cross oder mit großer Wahrscheinlichkeit rand) ist die FPCA viel schlechter als die ML-Verfahren. Insbesondere wird das Verfahren dadurch oft bei großem n schlechter.
- Die Schätzung von σ^2 ist allgemein schlecht.
- Das Verfahren hat große Schwierigkeiten mit kleinen Clustern.

- Um bei hoher Dimension zu guten Ergebnissen zu kommen, bräuchte man einen nicht vertretbaren Rechenaufwand (siehe Bemerkung 15.1). In der vorliegenden Version sind die Ergebnisse daher bei $p = 9$ und $p = 4$ bei kleinem n meistens schlecht.

Die hohe Zahl irrelevanter FPC in den Konstellationen mit $p = 4$ oder 9 könnte möglicherweise gesenkt werden, wenn die Mindestgröße für einen Cluster im Verfahren erhöht wird. Es ist nicht auszuschließen, daß in einer typischen Datenkonstellation im fünf- oder zehndimensionalen Raum tatsächlich 20 oder mehr Punktmengen kleiner Größe (zum Beispiel $< 3p$) so gut vom Rest der Daten getrennt sind, daß sie die Bezeichnung „Cluster" verdienen. Vergleichbare Effekte diskutiert Rousseeuw (1994) in einem Abschnitt über den „Fluch der Dimensionalität" („The curse of dimensionality"). Er schreibt dort über höherdimensionale Datensätze:
My interpretation of the „curse of dimensionality" is that several structures can exist simultaneously in the same dataset.

Die entsprechenden Ergebnisse der FPCA wären dann nicht unsinnig, da die FPCA anschaulich bedeutsame und nicht nur unbedingt modellseitig vorhandene Cluster finden soll.

Wenn eine der Modellvoraussetzungen aus Abschnitt 2 einigermaßen gesichert ist und man an einer guten Parameterschätzung interessiert ist, kann die FPCA mit den anderen Verfahren nicht konkurrieren. Wenn unsystematische Ausreißer auftreten, die Erfüllung der Modellvoraussetzungen sehr unklar ist oder man an einer genaueren Analyse des Datensatzes bei möglichst kleinem p interessiert ist, kann die FPCA wertvolle Informationen bringen. Möglicherweise ist es auch sinnvoll, die FPCA zur Ausreißeranalyse einzusetzen, wenn danach eines der anderen Verfahren angewendet werden soll.

17.2 Mischmodell-Maximum Likelihood

Vorteile:

- Das Verfahren liefert als einziges brauchbare Ergebnisse in der Konstellation 1rand3. Es ist zu vermuten, daß es bei Erfülltheit der Zuordnungsunabhängigkeit und Clustern, die sich stark überlappen, immer die besten Ergebnisse bringen wird.
- Von DeSarbo und Cron (1988) und Kiefer (1978) wird Konsistenz des Verfahrens im Falle bekannter Clusterzahl behauptet. Dieser Verdacht wird durch das Verhalten bei großem n in den Simulationen bestätigt; in den Fällen, wo die Parameterschätzung mit großem n schlechter wird (zum Beispiel lok), scheint es an der Schätzung der Clusterzahl zu liegen. Auch die Schätzung der Clusterzahl scheint aber meistens gegen den wahren Wert zu konvergieren.
- Der Effizienzverlust gegenüber dem KQ-Schätzer für die einzelnen Cluster ist in vielen Situationen sehr klein. Allerdings entsteht durch die Störskalenuntergrenze 0.001 bei den Clustern mit Störvarianz 0.001 ein beschönigtes Bild.

Nachteile:

- Das Verfahren ist häufig bei kleinem n schlecht.

- Das BIC überschätzt die Clusterzahl häufig.
- Die Schätzung der Zuordnungen der Punkte zu den Clustern ist den anderen Verfahren - insbesondere FPML - häufig unterlegen.
- Bei Verletzung der Zuordnungsunabhängigkeit verliert das Verfahren an Qualität.

Weiterhin ist MML bei großem n extrem langsam. Umgekehrt ist bei kleinem n eine mehrfache Wiederholung der Iteration zeitlich nicht problematisch und könnte zu Verbesserungen führen.

Als Fazit ist MML zu empfehlen, wenn die Modellvoraussetzungen stimmen, was man natürlich nie genau weiß. Weiter sollte der Datensatz nicht zu klein sein. Insbesondere ist das Verfahren anscheinend das einzige, das die Regressionsparameter vernünftig schätzt, wenn kein deutliches Muster vorhanden ist. Die Information, daß eine lineare Regression in Clustern sinnvoll ist, muß dann allerdings aus anderen Quellen kommen als aus der Analyse des Datensatzes allein.

In den anderen Situationen ist MML immer mindestens einem der anderen Verfahren deutlich unterlegen.

17.3 Fixed Partition Maximum Likelihood

Vorteile:

- Die Schätzung der Clusterzahl ist meistens hervorragend, neigt bei kleinem n eher zur Unterschätzung.
- Das Verfahren ist in den meisten der simulierten Konstellationen das beste, bei Verletzung der Zuordnungsunabhängigkeit und Abwesenheit von Ausreißern sogar praktisch immer.
- Auch bei Zuordnungsunabhängigkeit ist die Zuordnungsschätzung fast immer die beste.
- Die theoretische Inkonsistenz des Verfahrens (siehe Bemerkung 3.4) ist bei den hier verwendeten Stichprobenumfängen offenbar fast immer irrelevant.

Nachteile:

- In der Situation 1randl3 versagt FPML völlig. Vermutlich ist das Verfahren ungeeignet, wenn die Überschneidungen der Cluster zu groß sind.
- Das Verfahren hat die größten Probleme mit einer kleinen Zahl von Ausreißern.

Insgesamt macht FPML in den Simulationen also den besten Eindruck. Das liegt unter anderem an der Schwerpunktsetzung auf gut getrennte Cluster. Außerdem ist das Verfahren am schnellsten. Problematisch wäre es allerdings, wenn das Verfahren auf die meisten Arten von Verletzungen der Modellvoraussetzungen (zum Beispiel Ausreißer) empfindlich reagieren würde. Es wäre interessant, ob FPML eine größere Zahl von „modellfremden", d.h. nichtlinearen Ausreißern zu einem eigenen Cluster zusammenfassen würde und damit den Rest der Daten vernünftig behandeln könnte.

18 Schlußbetrachtung

18.1 Konsequenzen für die Anwendung

Angenommen, man steht vor einem Datenanalyse-Problem, in dem es darum geht, Cluster linearer Regression zu finden. Was soll man tun? Meine Arbeit trägt folgendes zur Beantwortung dieser Frage bei:

- Die Fixpunktclusteranalyse wurde eingeführt und theoretisch untersucht.
- Die ML-Schätzung für Modelle mit fester Zuordnung wurde auf den Regressionsfall übertragen. Ein Verfahren zur Schätzung der Clusterzahl wurde vorgeschlagen.
- Für die ML-Schätzung im Mischmodell wurde die Verwendung des Schwarz'schen Kriteriums zur Schätzung der Clusterzahl vorgeschlagen.
- Die Berechnung aller drei Verfahren wurde beschrieben und ausführlich diskutiert. In einer großen Simulation wurde ihr Verhalten in unterschiedlichen Datenkonstellationen untersucht.
- Unterschiedliche Möglichkeiten zur Modellierung wurden ausgeführt. Für alle Modelle wurden Bedingungen für die Identifizierbarkeit der Parameter hergeleitet.

Bevor man eines der hier untersuchten Verfahren anwendet, sollte überprüft werden, ob die Clusterbildung eventuell in einfacher Weise von der Zeit oder einem anderen eindimensionalen Regressor abhängt. In diesem Fall kann man häufig mit einem Verfahren für Wechselpunktprobleme bessere Resultate erreichen.

Die FPML-Schätzung taucht in der Literatur bisher im Lokationsfall nur selten, im Regressionsfall nie auf. Die theoretische Inkonsistenz wirkt offenbar abschreckend. Im Regressionsfall hat das Verfahren aber große Vorteile: Es ist deutlich schneller als die Konkurrenten, benötigt weniger restriktive Voraussetzungen als MML und schneidet in den Simulationen ausgezeichnet ab. Problematisch scheint jedoch die Anfälligkeit gegen Ausreißer zu sein.

Die Anwendung des MML-Verfahrens kann ich höchstens empfehlen, wenn es gute Gründe gibt, die Zuordnungsunabhängigkeit für erfüllt zu halten. Anderenfalls versagt es zwar nicht völlig, kann aber nicht mit FPML mithalten. Das gilt auch, wenn der Schwerpunkt auf der Zuordnungs- und nicht auf der Parameterschätzung für Regression und Störvarianz liegt.

Die FPCA halte ich für noch nicht ausgereift genug, um mit ihr alleine ein solches Datenanalyseproblem lösen zu wollen.[27] Das entscheidende Problem sind die Schwierigkeiten, Cluster zu finden, die deutlich weniger als die Hälfte der Daten enthalten. Auch der horrende Rechenaufwand bei hohen Dimensionen ist ein großer Nachteil. Die FPCA kann aber im Zusammenspiel mit anderen Verfahren gute Dienste leisten: Vor einer Analyse mit MML, FPML oder einem Wechselpunktverfahren kann die FPCA einzelne Ausreißer finden. Nachher kann sie auf einzelne Cluster separat angewendet werden, um

[27] In komplexen Datenanalysesituationen, zu denen die hier behandelten zählen, sollte man ohnehin mehrere Verfahren verwenden. Es gibt auch gute Gründe (Unrobustheit), den Ergebnissen isoliert angewendeter ML-Verfahren nicht zu trauen. Die Schwächen der FPCA in den simulierten Situationen mit $s > 2$ bzw. $p = 9$ sind aber von anderer Qualität.

zu testen, ob diese wirklich homogen sind. Im Zusammenhang mit einer robusten Regressionsschätzung kann getestet werden, ob Ausreißer vorhanden sind, die sich zu einem eigenen Cluster zusammenfügen lassen. Es kann überprüft werden, ob die „guten Daten" der robusten Regression homogen sind. In allen diesen Fällen können Teildatensätze ausgewählt werden, für die die Fixpunktclustereigenschaft gezielt überprüft werden kann oder die als Startpunkt des Algorithmus verwendet werden können. Zumindest kann die Anzahl der sinnvollen Startpunkte für den Algorithmus eingeschränkt werden. Es entfällt damit die Schwierigkeit, zufällig in den Teilmengen des gesamten Datensatzes herumsuchen zu müssen.

18.2 Ausblick

Die Idee der FPCA ist ausbaufähig. Im Regressionsfall wäre zu überprüfen, ob die Ersetzung des KQ-Schätzers durch robustere Schätzungen Verbesserungen bringt. Angenommen, ein Algorithmus analog zu dem aus Abschnitt 9 wäre auch mit einem robusten Schätzer konvergent. Für eine einzelne Iteration dieses Algorithmus ist dann eine Verlängerung der Rechenzeit zu befürchten, weil die Berechnung robuster Schätzer häufig aufwendig ist. Andererseits habe ich die Hoffnung, daß weniger Iterationen benötigt werden, um kleinere Cluster zu finden: Eventuell müssen nur noch knapp mehr als die Hälfte der Punkte des Iterationsstartes aus demselben Cluster sein, um den entsprechenden FPC zu iterieren. Gerade bei hoher Dimension könnte das viele Iterationsdurchläufe sparen (siehe die Diskussion in Bemerkung 15.1). Außerdem könnte sich die Anzahl der irrelevanten FPC verringern. Die Unrobustheit der KQ-Schätzung bewirkt zum Beispiel, daß fast immer der komplette Datensatz einen FPC bildet. Ich habe die KQ-Schätzung vor allem verwendet, weil sie theoretisch und numerisch am einfachsten handhabbar ist. Die Übertragung der Ergebnisse aus Abschnitt 9 und Teil III von KQ-FPC auf Fixpunktcluster, die zum Beispiel auf MM-Schätzern basieren, dürfte schwierig sein.

Weiterhin läßt sich die FPCA auf andere Problemstellungen übertragen. Abschnitt 7.2 steht stellvertretend für viele weitere Möglichkeiten.

Für KQ-FPC könnte nach Verfahren gesucht werden, relevante von irrelevanten Clustern zu unterscheiden. Man könnte nach Relevanzkriterien suchen, die zum Beispiel vom Verhältnis Varianz zu Clustergröße abhängen. Die Resultate aus Teil III könnten auf feste Regressoren (Fixed Partition Model) übertragen werden. Leider läßt meine Arbeit die Frage der Konsistenz der FPCV für die FPCI offen. Das größte Problem scheint mir hierbei zu sein, daß alle Existenzresultate für FPCI die Stetigkeit der Verteilungen voraussetzen. In alle Beweise geht entscheidend der Zwischenwert- bzw. Brouwersche Fixpunktsatz ein. Beide beruhen auf Stetigkeitseigenschaften. Um Aussagen über „benachbarte" empirische Verteilungen zu machen, bräuchte man vergleichbare Aussagen über Fixpunkte spezieller unstetiger Funktionen.

Weitere theoretische Anknüpfungspunkte ergeben sich in Bezug auf die MML- bzw. FPML-Schätzung. Für die MML-Schätzung ist das asymptotische Verhalten nicht ausreichend untersucht. FPML schätzt zwar inkonsistent; es wäre aber interessant, ob sich mit Hilfe einer Funktionalformulierung Schranken für die theoretische Verzerrung herleiten ließen. Es müßte also das FPML-Analogon zu den FPCI definiert und eine Theorie entsprechend Teil III dieser Arbeit entwickelt werden. Die theoretische Untermauerung der Clusterzahl-Schätzungen beider ML-Verfahren steht weiterhin aus.

Abbildungsverzeichnis

Symbolverzeichnis

Angegeben ist die Seite, auf der das Symbol zum ersten Mal definiert wird. Mit „(+)" gekennzeichnete Symbole und Symbole mit mehreren Seitenangaben haben nicht immer dieselbe Bedeutung, werden aber meistens in einem ähnlichen Zusammenhang verwendet. Nicht verzeichnet sind Symbole, die ihre Bedeutung nur in einer lokalen Umgebung (z.B. einem einzelnen Beweis) haben.

Index

Literatur

Akaike, H. (1974): A new look at the statistical identification model, *IEEE Transactions on Automatic Control* 19, S. 716-723

Bandelt, H.-J. und Dress, A. M. W. (1994): An order theoretic framework for overlapping clustering, *Discrete Mathematics* 136, S. 21-37

Banfield, J. D. und Raftery, A. E. (1993): Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics* 49, S. 803-821

Barnett, V. und Lewis, T. (1984): *Outliers in Statistical Data*, 2nd ed., Wiley, New York

Bauer, H. (1990): *Maß- und Integrationstheorie*, de Gruyter, Berlin

Billingsley, P. (1986): *Probability and Measure*, 2nd ed., Wiley, New York

Bock, H.-H. (1969): *The equivalence of two extremal problems and its application to the iterative classification of multivariate data*, Vortragsskript, Mathematisches Forschungsinstitut Oberwolfach

Bock, H.-H. (1996): Probability Models and Hypothesis Testing in Partitioning Cluster Analysis in Arabie, P., Hubert, L. J., und De Soete, G. (Hrsg.): *Clustering und Classification*, World Scientific Publications, New York, S. 377-453

Charles, C. (1979): Modeles lineaires locaux in Bochi, S. et alii (Hrsg.): *Optimisation en classification automatique*, INRIA, Le Chesnay, S. 367-428

Curcic, V. und Pierantoni, M. (1995): *The Analysis of Two-line Like Regression by Means of LMS*, Vortragsskript, ETH Zürich

Davies, P. L. (1995): Data features, *Statistica Neerlandica* 49, S. 185-245

Davies, P. L. und Gather, U. (1989): *The Identification of Multiple Outliers*, Forschungsbericht Nr. 89/1, Fachbereich Statistik, Universität Dortmund

Davies, P. L. und Gather, U. (1993): The Identification of Multiple Outliers with discussion, *Journal of the American Statistical Association* 88, S. 782-801

Day, N.E. (1969): Estimating the Components of a Mixture of Normal Distributions, *Biometrika* 56, S. 463-474

Dempster, A., Laird, N. M. und Rubin, D. B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm with discussion, *Journal of the Royal Statistical Society*, Series B, 39, S. 1-38

DeSarbo, W. S. und Cron, W. L. (1988): A Maximum Likelihood Methodology for Clusterwise Linear Regression, *Journal of Classification* 5, S. 249-282

Fahrmeir, L. und Hamerle, A. (Hrsg.) (1984): *Multivariate statistische Verfahren*, de Gruyter, Berlin

Gänssler, P. und Stute, W. (1977): *Wahrscheinlichkeitstheorie*, Springer, Berlin

Heuser, H. (1981): *Lehrbuch der Analysis, Teil 2* , Teubner, Stuttgart

Hinderer, K. (1970): *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*, Springer, Berlin

Hosmer, D. W. jr. (1974): Maximum Likelihood estimates of the parameters of a mixture of two regression lines , *Communications in Statistics* 3, S. 995-1006

Huang, W.-T. und Pao, K.-M.(1991): Minimum Distance Estimations in a Switching Regression Model, *Information and Management Sciences* 2, S. 119-128

Huber, P. J. (1981): *Robust Statistics*, Wiley, New York

Huskova, M. (1996): Estimation of a change in linear models, *Statistics and Probability Letters* 26, S. 13-24

Jacobson, H. I. (1969): The maximum variance of restricted unimodal distributions, *Annals of Mathematical Statistics* 40, S. 1746-1752

Jajuga, K. (1986): On pattern recognition methods in econometric regression model in Pau, L. F.: *Artificial Intelligence in Economics and Management*, Elsevier Science Publishers, Amsterdam, S. 167-171

Kiefer, N. M. (1978): Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model, *Econometrica* 46, S. 427-434

Krishnaiah, P. R. und Miao, B. Q. (1988): Review about estimation of change-points in Krishnaiah, P. R. und Rao, P. C.: *Handbook of Statistics Vol. 7*, Elsevier Science Publishers, Amsterdam

Leroux, B. G. (1992): Consistent estimation of a mixture distribution, *Annals of Statistics* 20, S. 1350-1360

Marriott, F. H. C. (1975): Separating mixtures of normal distributions, *Biometrics* 31, S. 767-769

Maturana, H. R. (1970): Neurophysiology of Cognition in P.L. Garvon (ed.): *Cognition: A Multiple View*, Spartan Books, New York, S. 3-23

Morgenthaler, S. (1990): Fitting Redescending M-Estimators in Regression in Lawrence, H. D. und Arthur, S. (Hrsg.): *Robust Regression*, Dekker, New York, S. 105-128

Odeh, R. E. und Evans, J. O. (1974): Algorithm AS 70: The Percentage Points of the Normal Distribution, *Applied Statistics* 23, S. 96-97

Piaget, J. (1975): *L'Equilibration des structures cognitives*, P.U.F., Paris

Plackett, R. L. (1950): Some theorems in Least Squares , *Biometrika* 37, S. 151-157

Prakasa Rao, B. L. S. (1992): *Identifiability in Stochastic Models: Characterizations of Probability Distributions*, Academic Press, Boston

Quandt, R. E. (1958): The estimation of the parameters of a linear regression system obeying two seperate regimes, *Journal of the American Statistical Association* 53, S. 873-880

Quandt, R. E. (1972): A New Approach to Estimating Switching Regressions, *Journal of the American Statistical Association* 67, S. 306-310

Quandt, R. E. und Ramsey, J. B. (1978): Estimating Mixtures of Normal Distributions and Switching Regressions with discussion, *Journal of the American Statistical Association* 73, S. 730-752

Rousseeuw, P. J. (1994): Unconventional features of positive-breakdown estimators, *Statistics and Probability Letters* 19, S. 417-431

Rousseeuw, P. J. und Leroy, A. M. (1988): *Robust Regression and Outlier Detection*, Wiley, New York

Rousseeuw, P. J. und Yohai, V. J. (1988): Robust Regression by means of S-estimators in Franke, J., Härdle, W. und Martin, R. D.: *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics 26, Springer, New York, S. 256-272

Schwarz, G. (1978): Estimating the dimension of a model, *Annals of Statistics* 6, S. 461-464

Scott, A. J. und Symons, M. J. (1971): Clustering methods based on likelihood ratio criteria, *Biometrics* 27, S. 387-397

Spaeth, H. (1979): Clusterwise Linear Regression , *Computing* 22, S. 367-373

Szekli, R. (1995): *Stochastic Ordering and Dependence in Applied Probability*, Lecture Notes in Statistics 97, Springer, New York

Teicher, H. (1961): Identifiability of Mixtures, *Annals of Mathematical Statistics* 34, S. 244-248

Titterington, D. M., Smith, A. F. M., Makov, U. E. (1985): *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester

von Förster, H. (1973): On Constructing a Reality in: W.F.E. Preiser (Hrsg.): *Environmental Design Research*, vol. 2, Stroudberg, S. 35-46

von Förster, H. (1976): Objects: Tokens for Eigen-Behaviors, *ASC Cybernetic Forum* 8, S. 91-96

von Förster, H. (1993): *Wissen und Gewissen* (hrsg. von S.J. Schmidt), Suhrkamp, Frankfurt/Main

Wedel, M. und Steenkamp, J.-B. E. M. (1991): A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation , *Journal of Marketing Research* 28, S. 385-396

Yakowitz, S. J. und Spragins, J. D. (1968): On the identifiability of finite mixtures, *Annals of Mathematical Statistics* 39, S. 209-214

Yao, Y.-C. (1988): Estimating the number of changepoints via Schwarz' criterion, *Statistics and Probability Letters* 6, S. 181-189

Yohai, V. J. (1988): High breakdown-point and high efficiency estimates for regression, *Annals of Statistics* 15, S. 642-665

Zelen, M. und Severo, N. C. (1964): Probability Functions in Abramowitz, M. und Stegun, I. A.: *Handbook of Mathematical Functions*, Dover Publications, New York, S. 925-996

Zusammenfassung

Eine lineare Regression kann durch eine Familie von Verteilungen $(P_{\beta,\sigma^2} : \beta \in I\!R^{p+1}, \sigma^2 \in I\!R^+)$ für $(x, y) \in I\!R^{p+1} \times I\!R$ modelliert werden, wobei $y = x'\beta + u$; u sei unabhängig von x und normal- oder zumindest symmetrisch um 0 verteilt mit Varianz σ^2.

In dieser Arbeit geht es um die Analyse von Datensätzen $(x_i, y_i) \in I\!R^{p+1} \times I\!R, i = 1, \ldots, n$. Eine lineare Regressions-Verteilung P_{β,σ^2} wird als Verteilung eines Clusters behandelt, d.h. lineare Regressionen mit unterschiedlichen Parametern $(\beta_i, \sigma_i^2), i = 1, \ldots, s$ sollen zur Modellierung unterschiedlicher Teile des Datensatzes adäquat sein. Es können auch Ausreißer unter den Daten sein, für die keine solche Verteilung angemessen ist.

Verschiedene Modelle für derartige Daten werden vorgestellt, insbesondere Mischmodelle der Form $\sum_{i=1}^{s} \epsilon_i P_{\beta_i,\sigma_i^2}$. Die Schätzung der Parameter (β_i, σ_i^2) mit Maximum Likelihood-Verfahren wird diskutiert. Für die Schätzung der Clusterzahl s werden neue Vorschläge gemacht.

Hinreichende Bedingungen für die Identifizierbarkeit der Parameter werden hergeleitet. In Situationen, in denen diese nicht erfüllt sind, werden einige Gegenbeispiele angegeben.

Ein neues Verfahren, die Fixpunktclusteranalyse (FPCA), wird eingeführt. Sie ermöglicht die Analyse von Datensätzen, die Ausreißer enthalten und bei denen die Anzahl der Cluster s unbekannt ist. Die FPCA basiert auf der Identifikation von Ausreißern und kann auch auf andere Clusterprobleme verallgemeinert werden. Ein Fixpunktcluster (FPC) ist eine Teilmenge des $I\!R^{p-1} \times I\!R$ und soll Punkte (x, y) umfassen, die zusammengehörig sind. Jeder FPC korrespondiert zu Parametern $(b, s^2) \in I\!R^{p+1} \times I\!R^+$, die als Schätzung der Regressionsparameter (β_i, σ_i^2) interpretiert werden können. FPC werden für Datensätze und Verteilungen definiert.

Ein konvergenter Algorithmus zur Berechnung von FPC in Datensätzen wird hergeleitet.

Verteilungen der Form $(1 - \epsilon)P_{\beta_0,\sigma_0^2} + \epsilon H^*$ werden betrachtet. Dabei wird P_{β_0,σ_0^2} als Verteilung eines Regressionsclusters interpretiert. H^* ist eine Verteilung auf $I\!R^{p+1} \times I\!R$, zum Beispiel eine Mischung weiterer P_{β,σ^2}. Unter verschiedenen Voraussetzungen an H^* und ϵ wird die Existenz von FPC gezeigt, deren Parameter in einer beschränkten Umgebung von (β_0, σ_0^2) liegen. Insbesondere existiert für homogene Regressionsverteilungen $(\epsilon = 0)$ genau ein FPC. Dieser hat die Parameter (β_0, σ_0^2).

In einer Simulationsstudie werden die FPCA und zwei Maximum Likelihood-Verfahren miteinander verglichen.

Christian Hennig - **Lebenslauf**

1966, 14.12. Geburt in Hamburg, Mutter Gertrud Hennig geb. Sick, Vater Martin Hennig, Pastor.

1973 Einschulung Grundschule Poßmoorweg, Hamburg

1977 Einschulung Heinrich-Hertz-Schule (kooperative Gesamtschule, Bereich Gymnasium), Hamburg

1986 Abitur mit Leistungsfächern Mathematik und Musik

1986 Beginn des Studiums der Mathematik an der Universität Hamburg

1988 Vordiplom mit Anwendungsfach Wirtschaftswissenschaften

1990 Anderthalb Jahre Gaststudium der Statistik an der Universität Dortmund

1993 Diplom in Mathematik, Spezialgebiet Statistik mit Note „sehr gut", Anwendungsfach Soziologie. Diplomarbeit: *Eine Einflußanalyse für S-, SM- und Sμ-Schätzer in Regressionsmodellen*, betreut von Professor Behnen am Institut für Stochastik der Universität Hamburg.

1993 Antritt meiner Stelle als Wissenschaftlicher Mitarbeiter im Institut für Mathematische Stochastik der Universität Hamburg, Forschung auf den Gebieten Robuste Statistik und Clusteranalyse Linearer Regression

1994 Leitung der Orientierungseinheit für Studienanfänger/innen

1995 Konzeption und Leitung der Arbeitsgemeinschaft Datenanalyse, Veröffentlichung Efficient high-breakdown-point estimaturs in robust regression: Which function to choose?, *Statistics and Decisions 13*, 221-241

1997 Fertigstellung meiner Dissertation über *Datenanalyse mit Modellen für Cluster Linearer Regression*, betreut von Professor Behnen, Zweitgutachter Professor Pfeifer

1997 Vorlesung über „Robuste Mathematische Statistik" an der Universität Hamburg